UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Information Retrieval using Machine Learning for Database Curation

Sofia Pinheiro Rodrigues de Jesus

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Prof. Doutor Francisco José Moreira Couto

2019

Resumo

Em 2016, a Agência Internacional de Pesquisa sobre o Cancro da Organização Mundial de Saúde lançou a primeira base de dados de biomarcadores de exposição, chamada Exposome-Explorer. Para construir a base de dados, mais de 8500 citações foram manualmente analisadas, mas apenas 480 foram consideradas relevantes e usadas para extrair informação para integrar a base de dados.

Curar manualmente uma base de dados é uma tarefa demorada e que requer especialistas capazes de recolher e analisar dados que se encontram espalhados por milhões de artigos. Esta tese propõe o uso de técnicas de Recuperação de Informação com uma abordagem de aprendizagem supervisionada para classificar automaticamente artigos como relevantes ou irrelevantes para auxiliar o processo de criação e atualização da Exposome-Explorer. Esta abordagem restringe a literatura a um conjunto de publicações relevantes sobre biomarcadores de exposição de uma maneira eficiente, reduzindo o tempo e esforço necessários para identificar documentos relevantes. Além disso, as *queries* originais usadas pelos curadores para pesquisar sobre literatura de biomarcadores de exposição foram melhoradas para incluir alguns artigos relevantes que anteriormente não estavam a ser encontrados.

Os dados manualmente recolhidos da Exposome-Explorer, foram usados para treinar e testar os modelos de aprendizagem automática (classificadores). Vários parâmetros e seis algoritmos diferentes foram avaliados para averiguar quais previam melhor a relevância de um artigo com base no título, resumo ou metadados. O melhor classificador foi construído com o algoritmo SVM e treinado com os resumos dos artigos, obtendo um *recall* de 85.8%. Este classificador reduz o número de citações sobre biomarcadores dietéticos a serem manualmente analisadas pelos curadores em quase 88%, classificando apenas incorretamente 14.2% dos artigos relevantes. Esta metodologia também pode ser aplicada a outros dados de biomarcadores ou ser adaptada para auxiliar o processo de criação manual de outras bases de dados químicas ou de doenças.

Palavras Chave: Aprendizagem Automática, Prospeção de Texto, Recuperação de Informação, Biomarcadores de exposição, Curação de base de dados

Abstract

In 2016, the International Agency for Research on Cancer, part of the World Health Organization, released the Exposome-Explorer, the first database dedicated to biomarkers of exposure for environmental risk factors for diseases. To build the database more than 8500 citations were manually screened, but only 480 were used to extract information to develop the database. Manually curating a database is time-consuming and often requires domain experts to gather relevant data scattered throughout millions of articles.

This thesis proposes using Information Retrieval techniques with a supervised learning approach to automatically classify articles as relevant or irrelevant to assist the curation process of the Exposome-Explorer database. This approach narrows down the literature to a set of relevant publications about biomarkers of exposure in a time-efficient manner, hence reducing the effort necessary to identify relevant papers for the database. In addition, the original queries used by the curators to search for literature regarding biomarkers of exposure were improved to include some relevant articles that were not being targeted before.

The manually selected corpus of scientific publications used in the Exposome-Explorer was used as a training and testing set for the machine learning models (classifiers). Several parameters and six different algorithms (Decision Tree, Logistic Regression, Naïve Bayes, Neural Network, Random Forest and Support Vector Machine) were evaluated to predict an article's relevance based on its title, abstract or metadata. The best classifier was built with the Support Vector Machine algorithm using the abstracts' set, achieving a recall of 85.8%. This classifier reduces the number of dietary biomarker citations to be manually screened by the database curators by nearly 88%, while only misclassifying 14.2% of the relevant articles. This methodology can also be applied to similar biomarkers datasets or be adapted to assist the manual curation process of similar chemical or diseases databases.

Keywords: Machine Learning, Text Mining, Information Retrieval, Biomarkers of exposure, Database curation

Resumo Alargado

Biomarcadores são parâmetros biológicos objetivamente medidos no corpo que funcionam como indicadores de condições biológicas normais, patológicas, estilos de vida ambientais ou como resposta a intervenções terapêuticas. Caracterizar a relação entre biomarcadores e os possíveis resultados biológicos é crucial para prever corretamente respostas clínicas, diagnosticar pacientes, identificar riscos, avaliar a exposição a agentes patogénicos e melhorar a eficiência dos ensaios clínicos. Biomarcadores de exposição são substâncias exógenas, os seus metabolitos ou produtos de interação com moléculas, que refletem a exposição de um indivíduo a um fator ambiental. Estes compostos podem entrar em contato com organismos através da absorção, inalação ou ingestão, sendo de seguida metabolizados, armazenados ou eliminados. Através da análise de amostras biológicas, como sangue ou urina, ou da medição de concentrações é possível detetar essa exposição.

Em 2016, a Agência Internacional de Pesquisa sobre o Cancro da Organização Mundial de Saúde, lançou a primeira base de dados de biomarcadores de exposição, chamada Exposome-Explorer. A base de dados contém informações detalhadas sobre a natureza de 692 biomarcadores dietéticos e poluentes extraídas de 480 publicações. Um total de 10 510 valores de concentração medidos no sangue, urina e noutras amostras biológicas foram registados. Os detalhes sobre as populações em que os biomarcadores foram medidos e as técnicas analíticas utilizadas na medição também estão armazenados na base de dados. Para além disso, contém ainda 8034 valores de correlação entre os níveis de biomarcadores dietéticos e a ingestão de alimentos e 536 valores de reprodutibilidade biológica ao longo do tempo. Para encontrar artigos sobre biomarcadores de exposição para desenvolver a Exposome-Explorer, foi realizada uma pesquisa na Web of Science (WOS) usando *queries* com palavras-chave específicas associadas a biomarcadores dietéticos, de poluição e valores de reprodutibilidade. Mais de 8500 citações de 1975 até 2014 foram manualmente analisadas, mas apenas 480 foram consideradas relevantes e usadas para extrair informação para construir a base de dados. Reunir dados relevantes espalhados por milhões de artigos e repositórios é uma tarefa demorada e que requer especialistas capazes de recolher e analisar todos estes dados. Portanto, este método não é a solução mais eficiente para encontrar novas informações nem para manter a base de dados atualizada.

O número de novas publicações científicas sobre biomarcadores adicionadas à WOS e ao PubMed tem crescido em milhares de artigos a cada ano que passa. Profissionais de saúde, como médicos ou investigadores, precisam constantemente de ter acesso a informações atualizadas e dependem destes recursos online para encontrarem dados essenciais que apoiam diariamente o seu trabalho. Reunir dados relevantes espalhados por milhões de documentos com texto não estruturado de repositórios biomédicos torna muito difícil encontrar rapidamente informações cruciais e requer profissionais especializados para examinar manualmente a literatura. Há uma necessidade crescente de automatizar a identificação e priorização de artigos para serem manualmente analisados. A Recuperação de Informação (IR) tem como objetivo automatizar o

processo de restringir uma coleção de documentos a um conjunto de documentos relevantes que contêm informações de interesse para um assunto específico. Essa tarefa pode seguir uma abordagem de Aprendizagem Automática (ML), que usa algoritmos para permitir que os computadores aprendam e melhorem automaticamente o desempenho de uma tarefa específica, sem precisarem de ser explicitamente programados para cada cenário possível. Em problemas como a criação e atualização de bases de dados, os métodos de classificação supervisionados são geralmente preferidos. Estes métodos usam dados previamente classificados para treinar um classificador de ML, que usa um conjunto de características dos dados para aprender e, posteriormente, para prever, a classe a que pertence cada elemento dos dados. Assim, IR com ML pode auxiliar o processo de criação e atualização de bases de dados biomédicas, como a Exposome-Explorer, identificando automaticamente artigos relevantes numa grande coleção de documentos biomédicos.

Esta dissertação visa reduzir o tempo, o esforço e os recursos necessários para manter a Exposome-Explorer atualizada através do uso de técnicas de IR. Os dois objetivos principais desta tese são: usar uma abordagem de aprendizagem supervisionada para melhorar a tarefa de IR de forma a auxiliar o processo de manutenção da Exposome-Explorer e melhorar as *queries* originais usadas pelos curadores da base de dados para pesquisar artigos na WOS. Todo o trabalho foi desenvolvido utilizando dados fornecidos pelos curadores da Exposome-Explorer, nomeadamente, as *queries* utilizadas para pesquisar literatura sobre biomarcadores de exposição na WOS, os resultados dessa pesquisa e os 480 artigos relevantes, selecionados desses resultados, usados para construir a base de dados.

Para atingir o primeiro objetivo, o PubMed foi utilizado para extrair os títulos, resumos e metadados (revista, ano, número de citações e autores) das citações presentes nos resultados da pesquisa da WOS. Depois dos dados serem recolhidos, foram pré-processados para depois conseguirem ser utilizados pelos modelos de ML. Como os resultados preliminares com os dados de todos os tipos de biomarcadores de exposição (dietéticos, poluentes e valores de reprodutibilidade) não foram muito altos devido à baixa qualidade dos dados, estes foram restringidos a apenas biomarcadores dietéticos. Com este subconjunto de dados foram criados 2880 classificadores de ML com diferentes combinações de parâmetros e algoritmos para avaliar quais previam melhor a relevância dos artigos com base no título, resumo, metadados ou título + metadados. O modelo com o *recall* mais elevado (85.8%) foi construído com o algoritmo SVM usando os resumos para prever a relevância de um artigo. Esse classificador reduziu o número de citações sobre biomarcadores dietéticos a serem analisados manualmente pelos curadores da base de dados em quase 88%, enquanto apenas classificaram erradamente 14.2% dos artigos relevantes.

Para confirmar que esta metodologia também pode ser aplicada a outros dados de biomarcadores semelhantes ou ser adaptada para auxiliar o processo de criação e atualização manual de outras bases de dados biomédicas, foram utilizados os dados do CIViCmine. A metodologia foi adaptada para prever a relevância de frases relacionadas com biomarcadores associados a genes, medicamentos e tipos de cancro. O melhor *recall* foi de 74.5% com o algoritmo SVM. Este classificador permitiu reduzir o tempo necessário para encontrar 74.5% das frases relevantes em 83.6%, com uma perda de 25.5% das relevantes.

Para atingir o segundo objetivo, as *queries* originais foram melhoradas para incluir os 79 artigos que foram utilizados para extrair informações sobre biomarcadores de exposição para a Exposome-Explorer, mas que não foram encontrados nos resultados de pesquisa da WOS. No geral, as melhorias das *queries* e a nova *query* desenvolvida para a poluição geral aumentaram o número total de artigos em 2231, permitindo que 39 destes artigos relevantes fossem encontrados. Não foi possível incluir os restantes 40 artigos nos resultados das pesquisas, ou porque não estavam na WOS ou porque o custo de os incluir era maior que o benefício. As alterações permitiram que as *queries* fossem mais personalizadas para as necessidades dos curadores, uma vez que foram modificadas para incluir termos específicos que estavam presentes em exemplos reais de artigos relevantes escolhidos manualmente pelos curadores, o que deve aumentar a percentagem de artigos relevantes encontrados nos resultados de pesquisa. No entanto, não foi possível quantificar o ganho de alterar as *queries*, pois os curadores da base de dados teriam que avaliar manualmente a relevância de cada um dos novos artigos.

Para aplicar esta metodologia ao processo de construção da base de dados, a tarefa de IR passaria a ter duas etapas. Na primeira, os artigos continuariam a ser pesquisados na WOS, mas com as *queries* melhoradas, para restringir os resultados a publicações específicas a biomarcadores de exposição. De seguida, o classificador seria usado para restringir ainda mais o conjunto de artigos, classificando-os como relevantes ou não para a Exposome-Explorer. Ainda seria necessário analisar manualmente as publicações, porém num conjunto de artigos muito mais restrito.

As principais contribuições deste trabalho são: um conjunto de dados, adaptado da Exposome-Explorer, com títulos, resumos e metadados de 7083 artigos científicos, classificados como relevantes ou irrelevantes de acordo com as informações que possuem sobre biomarcadores de exposição; uma metodologia para melhorar a tarefa de IR com uma abordagem de aprendizagem supervisionada para classificar artigos com base em seus resumos, títulos ou metadados (disponível no GitHub: `https://github.com/lasigeBioTM/BLiR`); um artigo submetido baseado no Capítulo 3.

Acknowledgements

First, I want to thank my advisor, Professor Francisco Couto, for guiding me through this year and for always being available to help me.

Secondly, I want to thank FCT for the fellowship that financially supported me during the thesis and for supporting this work through funding of DeST: Deep Semantic Tagger project, ref. PTDC/CCI-BIO/28685/2017 (`http://dest.rd.ciencias.ulisboa.pt/`), and LASIGE Research Unit, ref. UID/CEC/00408/2019.

I want to give a special thanks to André Lamúrias, who has always been available to help me, give me feedback and discuss ideas. Thank you for reading my work countless times and for always making time to help me.

In addition, I want to thank Reza Salek and Vanessa Neveu for kindly providing the data that made this work possible and for writing an article with me.

I also want to thank my friends for being with me all the way through. This year was so amazing because we got to do it together.

Last but not least, I want to thank my family, especially my parents and my sister, for inspiring me and giving me all the support I need to achieve my goals and Pedro Guerreiro for being my rock and motivating me every single day.

Contents

List of Figures

List of Tables

Acronyms

API Application Programming Interface.

CIViC Clinical Interpretation of Variants in Cancer.

DBP Disinfection Byproducts.

DOI Digital Object Identifier.

DT Decision Tree.

ER Entity Recognition.

FN False Negatives.

FP False Positives.

HCA Heterocyclic amines.

HCC Hepatocellular Carcinoma.

IR Information Retrieval.

LMT Logistic Model Trees.

LR Logistic Regression.

ML Machine Learning.

NB Naïve Bayes.

NCBI National Center for Biotechnology Information.

NLP Natural Language Processing.

NLTK Natural Language Toolkit.

NN Neural Networks.

PAH Polycyclic aromatic hydrocarbons.

PBB Polybrominated biphenyls.

PBDE Polybrominated diphenyl ethers.

PCB Polychlorinated biphenyls.

PCDD/F Polychlorodibenzo-p-dioxins and Polychlorodibenzo-furans.

PMID PubMed ID.

RE Relationship extraction.

RF Random Forest.

SVM Support Vector Machine.

TFIDF Term Frequency-Inverse Document Frequency.

TM Text Mining.

TN True Negatives.

TP True Positives.

TSV Tab separated values.

WOS Web of Science.

Chapter 1

Introduction

Biomarkers are biological parameters objectively measured in the body as indicators of normal biological conditions, environmental lifestyles, pathological conditions, or responses to therapeutic interventions (Strimbu & Tavel, 2010). They can be chemicals, metabolites, enzymes and other biochemical substances, like products of an interaction between a compound and a target molecule or cell. Characterizing the relationship between biomarkers and the possible biological outcomes is crucial to correctly predict clinical responses, screen, monitor and diagnose patients and to improve the efficiency of clinical trials. They also play a significant role in risk assessment, as they allow to identify hazards, assess exposure and to associate responses with the probability of a disease outcome.

Biomarkers can be classified as biomarkers of effect, susceptibility and exposure (WHO, 1993):

- **biomarker of effect:** a quantitative biochemical, physiological or behavioural change an individual can undergo in consequence of a health issue or illness;
- **biomarker of susceptibility:** an indicator of an intrinsic or acquired characteristic of an individual that determines how they respond to the exposure to an exogenous substance;
- **biomarker of exposure:** an exogenous substance, its metabolite or its products of interaction with target molecules or cells that reflects the exposure of an individual to an environmental factor (such as diet, pollutants and infectious agents) known to affect the etiology of diseases. Compounds can get in contact with living organisms through absorption, inhalation or ingestion and then are either metabolized, stored or eliminated. By analysing biospecimens, such as blood or urine, or by measuring concentrations and characterizing the exogenous substance, its metabolites or its products of interaction with target molecules, it is possible the detect this exposure. This work will focus specifically on this type of biomarkers.

Exposome-Explorer (`http://exposome-explorer.iarc.fr/`) is a manually curated database for biomarkers of exposure to environmental risk factors developed by the International Agency for Research on Cancer, part of the World Health Organization (Neveu *et al.*, 2016). It contains detailed information about the nature of 692 dietary and pollutant biomarkers from 480 peer-reviewed publications. A total of 10 510 concentration values measured in blood, urine, and other biospecimens have been registered. The details about the populations and subjects in which biomarkers have been measured and the analytical techniques used for measurement, are

also stored in the database. In addition, it contains 8034 correlation values between dietary biomarker levels and food intake and 536 values of biological reproducibility over time.

1.1 Motivation

To develop the Exposome-Explorer database, biomedical literature was searched on the Web of Science (WOS), using queries with specific keywords associated with dietary, pollutant and reproducibility biomarkers and biospecimens. There were several requirements for a scientific paper to be considered for the database, for example, it had to be peer-reviewed, describe original work with biomarker measurements in human observational studies and be available online. Citations from the articles meeting the requirements were downloaded in the BibTeX format and managed with Bib-Desk. More than 8500 citations from 1975 until 2014 were manually screened to identify information about biomarkers of exposure, however, only 480 of them were included in the Exposome-Explorer database after being manually analysed and annotated. This method is not the most viable solution to either collect new data nor to keep the database updated.

The number of new scientific publications about biomarkers added to both PubMed and WOS has been growing by thousands each year (Figure 1.1). Health and biomedical professionals, such as doctors or researchers, are in constant need to have access to up-to-date information and rely on many online data resources to provide them with essential information to support their work each day. Gathering relevant data scattered through millions of documents with unstructured text from biomedical repositories makes it very difficult to find crucial information quickly and requires specialized professionals to manually screen the literature.

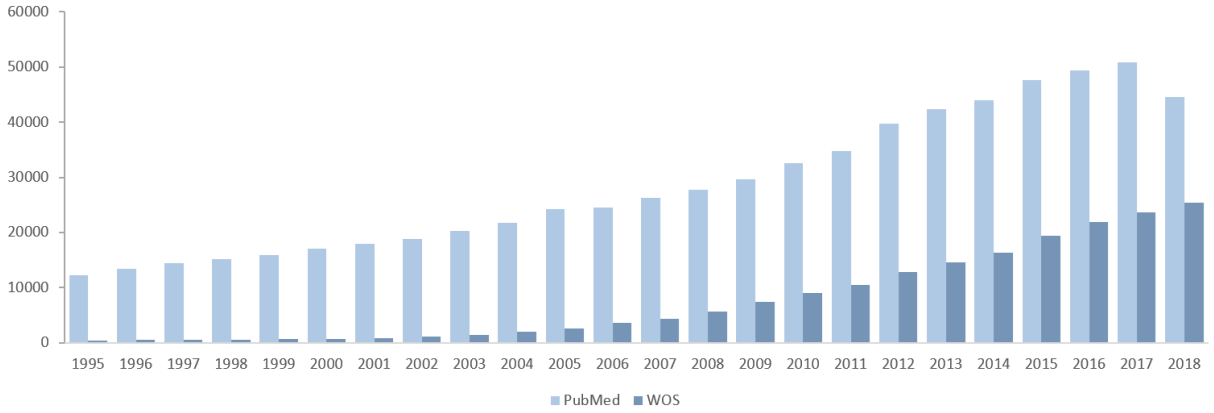

Figure 1.1: Number of new articles related to biomarkers added each year (from 1995 until 2019) on WOS and on PubMed.

There is an increasing need to automate the identification and prioritization of articles for manual curation. Information Retrieval (IR) is a task that automates the process of narrowing down a collection of documents to a set of relevant ones that contain information of interest for a specific subject. The IR task can be supported with Text Mining (TM) techniques and Machine Learning (ML), which uses algorithms to allow computer systems to automatically learn and improve the performance of a specific task, without having to be explicitly programmed for each possible outcome. There are several advantages of using ML in TM, for example, its accuracy is, in some cases, comparable to the one achieved by human experts, however, it is more cost and

time efficient (Sebastiani, 2002); on top of that, ML also allows to find unknown patterns in the data that otherwise could be missed (Witten *et al.*, 2016).

In problems such as database curation, supervised classification methods are commonly preferred. These methods use labelled data to train a ML classifier, which uses a set of features to learn, and later to predict, the class each element of the data belongs to. When classifying articles, there are usually two possible classes: "relevant" or "irrelevant", which is referred to as binary classification.

IR using ML can assist the manual curation process for several biomedical databases, such as Exposome-Explorer, by automatically identifying relevant articles among a large collection of biomedical documents.

1.2 Objectives

This work aims to reduce the time, effort and resources necessary to keep the Exposome-Explorer database updated as the number of articles increases. There are two main objectives for this thesis:

1. perform IR with a supervised learning approach to automatically retrieve and classify articles, based on their abstract, title or metadata, as relevant or irrelevant for the Exposome-Explorer database;

2. improve the original queries used by the database curators to search for articles on the WOS.

1.3 Methodology

The Exposome-Explorer database curators provided the search queries, the results from the WOS search and the 480 relevant articles used to construct the database.

The first part of the work focused on using the results from the queries and the list of relevant articles to train ML models to retrieve a subset of relevant articles about biomarkers of exposure from the literature. First, PubMed was used to gather all the necessary information about each article present in the WOS search results, namely the abstract, title and metadata (journal, date, number of citations and authors). Then, the text data was preprocessed into numerical data and labels (0 for irrelevant and 1 for relevant) were assigned to each element, based on whether they had been used to extract relevant information for the Exposome-Explorer database. The data was then separated into a training and testing set. The training set was used to build the ML models, called classifiers. Several parameters and six different algorithms (Decision Tree, Logistic Regression, Naïve Bayes, Neural Network, Random Forest and Support Vector Machine) were tested to predict an article's relevance based on its title, abstract or metadata. The testing set was used to compute the precision, recall and F-score of each classifier, by comparing the predictions to the real and known values. These metrics allowed to evaluate the performance of the classifiers. When given new data, the ML models should be able to predict if an article is relevant or not for the Exposome-Explorer the database.

The second part of the work aimed to improve the original queries used to search for literature regarding biomarkers of exposure on the WOS. The database curators added a few articles that they considered relevant but that were not retrieved automatically with the queries. Each article

was individually analysed to assess which improvements could be made to the original queries in order to included them and other similar articles in the results.

1.4 Contributions

The main contributions of this work can be enumerated as follows:

1. dataset with titles, abstracts and metadata from 7083 scientific papers, classified as relevant or irrelevant according to the information they hold about biomarkers of exposure, adapted from Exposome-Explorer;
2. IR-based methodology using a supervised learning approach to classify articles based on their abstracts, titles and metadata. This methodology can be applied to other biomarkers' datasets or be adapted to assist the curation of other databases in the areas of biology and medicine. Available on GitHub: `https://github.com/lasigeBioTM/BLiR`;
3. Article submitted: Jesus, S., Lamurias, A., Neveu, V., Salek, R. M., & Couto, F. M. Information Retrieval using Machine Learning for biomarker curation in the Exposome-Explorer. This article is based on Chapter 3.

1.5 Document Structure

The document structure is as follows:

- **Chapter 2:** provides the context needed to understand the work, such as concepts, description of similar works about information retrieval in the biomedical field and the resources used;
- **Chapter 3:** presents a methodology that uses machine learning to automatically classify articles as relevant or irrelevant to the Exposome-Explorer database;
- **Chapter 4:** focuses on improving the original queries used to search for articles about biomarkers of exposure on the WOS;
- **Chapter 5:** presents the main conclusions of this work and ideas for future work.

Chapter 2

Related Work

2.1 Background

2.1.1 Text Mining

Hearst (2003) has defined TM as "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. The goal is to discover heretofore unknown information, something that no one yet knows and so could not have yet written down." In TM, the patterns are extracted from unstructured natural language text and the extracted information links together to form new facts or hypothesis that can be validated through experiments. TM techniques can be used to address several tasks (Figure 2.1), such as:

- Information Retrieval (IR): automatically extracting resources that are relevant to a user from a large collection of documents, such as scientific papers. Querying databases with a set of keywords allows the user to obtain literature that fits their purpose. However, it might be necessary to narrow down that set of relevant documents even further;
- Entity Recognition (ER): clear identification of an entity in free text. That entity has a specific class assigned, depending on the concept it is been referred to in the document. It also allows an entity to be recognized by synonyms, for example, P53 and PT53 refer to the same gene;
- Relationship extraction (RE): identification and extraction of entities and their relations from a text.

Figure 2.1: Text Mining tasks: Information Retrieval, Entity Recognition and Information Extraction

2.1.1.1 Natural Language Processing

In TM, information is extracted from documents containing natural language. Natural Language Processing (NLP) (Manning *et al.*, 1999) is a sub-field of computer science, artificial intelligence and linguistics which aims to automatically process and analyse natural language. The following NLP techniques are commonly used in TM systems (Jurafsky & Martin, 2014):

- **Tokenization:** separate the text into units (tokens). Tokenization can occur at different levels, depending on what interests the user, for example, a text can be separated into chapters, paragraphs, sentences or words. The last two levels are the most common ones;
- **Stemming:** reduce all words with the same root to a common form (stem), usually by removing the suffixes and prefixes. The resulting stem is not necessarily a real word;
- **Stop Word Removal:** remove words that do not enrich the content of documents (such as "and", "or", "the", "but"). These words are as likely to appear in relevant documents as they are in irrelevant ones;
- **n-grams:** a contiguous sequence of n items from a given sample of text or speech. For example, for $n = 2$, the features from the sentence *"Determining thiocyanate serum levels"*, are combined into three n-grams: *"Determining thiocyanate"*, *"thiocyanate serum"* and *"serum levels"*;
- **Bag-of-Words:** a way to represent features extracted from text data for ML models. All instances of the same word that occur in a document are added to the same "bag", disregarding grammar and order;
- **Term Frequency-Inverse Document Frequency (TFIDF):** is a weight that reflects how important a word is to a document in a collection of documents. A specific word is considered more relevant as the number of times it appears in the same document increases (TF), however, if it also appears frequently in other documents from the same collection, it means its just a frequent word, with no relevant meaning (IDF).

2.1.1.2 Machine Learning

ML (Bishop, 2006) uses algorithms to allow computer systems to automatically learn and improve the performance of a specific task, without having to be explicitly programmed for each possible outcome. ML uses training data as input and through statistical analysis of the observed features, the created model is able to make inferences on new data. Once the model is created, the testing set can be used to evaluate its performance.

There are two main types of learning approaches, depending on the training data provided to train the model:

- **Supervised learning:** each element of the training set has known labels, which are used to train the model. Two types of supervised learning are:
  - **Classification:** when the output variable is a discrete value, like a category, for example, predicting if a book is a "romance", a "thriller" or "science fiction";
  - **Regression:** when the output variable is a continuous value, for example, predicting house prices;

- **Unsupervised learning:** the model is trained without a labelled training set. Two types of unsupervised learning are:
  - **Clustering:** group items with previously unknown similar characteristics;
  - **Dimensionality Reduction:** reduce the number of features that characterize an item to a set of main ones.

In IR, relevant documents are usually retrieved with queries, where documents are ranked according to some criterion. However, many relevant articles can be missed because they use terms that are not targeted in the queries. To overcome this problem, the queries need to be redefined over time, which increases complexity. Implementing a classification approach in IR can overcome this issue, as classes are usually more general than the terms targeted in the queries (Frakes & Baeza-Yates, 1992). This work uses a supervised binary classification approach, as the labels of each element of the dataset are known and have two possible classes: "irrelevant" (0) or "relevant" (1). The ML models created are called classifiers. Some commonly used algorithms with a supervised classification approach are: Decision Tree (DT), Logistic Regression (LR), Naïve Bayes (NB), Neural Networks (NN), Random Forest (RF) and Support Vector Machine (SVM).

Decision Tree (Quinlan, 1986) uses a tree-like model to make predictions. The tree begins at the root, then it has nodes that represent each feature, with a branch for each possible observation. The tree ends with leaves featuring the classes. The classification of an item starts at the root of the tree, where a question is made for the corresponding feature, for example: "Was the article published before the year 2010?", the branch followed is the one with the observed outcome. The process is repeated for each following node until a leaf is encountered, in which the respective class is assigned to the item.

Logistic Regression (Dietz *et al.*, 2002) is a mathematical modeling approach that can be used to describe the relationship between several independent variables to a binary dependent variable (contains exactly two possible values: 0 or 1). The LR model is based on the logistic function, an "S" shaped function that ranges from 0 to 1. The model is designed to make a class prediction based on a probability: 1 if that probability is >50% and 0 otherwise.

Naïve Bayes (Singh & Husain, 2014) is a simple probabilistic classifier based on Bayes' theorem with strong (Naïve) independence assumption. It assumes all features are independent and therefore they all contribute the same to the probability of an item being classified in a specific class.

Neural Networks (Fausett *et al.*, 1994) consist of several neurons organized in layers (input, hidden and output layers). Neurons of each layer are connected with an associated weight. Each neuron also has a bias value. A neuron takes the weights and bias as inputs, applies a nonlinear activation function and passes down the results between layers until the exit node, which produces a class. Each element of the training set is classified one by one. By comparing the predicted result to the actual value, the error is calculated and the weights are adjusted according to the learning rate (backpropagation).

Random Forest (Breiman, 2001) is a classifier consisting of a combination of decision trees, each tree is constructed using a different randomly selected sample of the same dataset. All trees in the forest are independent and have the same distribution. In standard trees, each node is split using the most important feature, but in a random forest, each node is split using the best feature among a random subset of features. It depends on two parameters: the number of variables in the random subset at each node and the number of trees in the forest. When given new data, its class is predicted by aggregating the predictions of all the trees in the forest.

Support Vector Machine (Cristianini *et al.*, 2000) represents data as points and each point is a n-dimensional vector (n is the number of features). This algorithm aims to find a hyperplane that maximizes the margin that separates data points with the same class. Hyperplanes are affine subspaces of dimension n-1. When given new data, the predicted class depends upon which side of the hyperplane that data point falls into.

2.1.1.3 Performance assessment

When developing a TM system, like a classifier, it is important to have metrics that can evaluate the performance of that tool and measure how relevant it is. In a supervised learning approach, the previously known labels attributed to each item are considered the gold-standard (an annotated corpus with labels corresponding to the actual results). In a binary classification problem, the classifier outputs a predicted label for the testing set, that can be either positive (1) or negative (0), depending if the desired criteria was verified or not. By comparing the items' predicted labels to the gold standard it is possible to separate them into one of four categories:

- True Positives (TP): documents correctly labelled as positive;
- False Positives (FP): documents incorrectly labelled as positive;
- True Negatives (TN): documents correctly labelled as negative;
- False Negatives (FN): documents incorrectly labelled as negative.

The confusion matrix (Table 2.1) is a table that allows to visualize this categorization.

Table 2.1: Confusion matrix

		Actural Values (gold-standard)	
		Positive (1)	Negative (0)
Predicted	Positive (1)	TP	FP
Values	Negative (0)	FN	TN

For example, the aim of a classifier is to predict if an article is relevant or not for a database. If the predicted label is 1, it means the article is relevant because the condition (being relevant) was verified. If the true label of that article is also 1, then it is considered a true positive, because the gold-standard and the predicted value are both positives.

Precision and Recall are two commonly used metrics that assess the performance of text-mining tools, by measuring the quality of the results in terms of relevancy. Precision (P) is the proportion of true positives items over all the items the system has labelled as positive. Recall (R) is the proportion of true positives items over all the items that should have been labelled as positive.

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F = \frac{2PR}{P + R}$$

The F-score is a measure between 0 and 1 that combines both precision and recall. Higher values of this metric indicate that the system classifies most items in the correct category, therefore having low numbers of FP and FN.

2.2 State of the Art

This thesis focuses on the IR task applied to the biomarker of exposure domain. Table 2.2 shows similar works developed in the area. Most studies related to biomarkers concentrate in the ER or RE tasks and none of them is specific to biomarkers of exposure. There are several papers that use ML to perform IR, but most target other biomedical fields.

Table 2.2: Summary of the studies developed in TM for biomedical literature and for biomarker literature.

		Domain	
		Biomarker literature	Biomedical literature
TM task	IR	Younesi *et al.* (2012)	BioCreAtIvE challenges; Almeida *et al.* (2014); Almeida *et al.* (2016); Krallinger *et al.* (2014)
	ER/RE	Lever *et al.* (2018); Chang *et al.* (2017); Bravo *et al.* (2014)	out of scope

2.2.1 IR in biomedical literature

The BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) challenge (Hirschman *et al.*, 2005) aims to evaluate text mining and information extraction systems applied to the biological domain. This initiative allows the community to assess current scientific progress. There have been a few BioCreative challenges dedicated to bioliterature text classification, such as BC Workshop '12 Track I- Triage (Lu & Hirschman, 2012), BioCreative IV Interactive Task (IAT) for Biocurators (Matis-Mitchell *et al.*, 2013), BioCreative V Interactive Task (IAT) for Biocurators (Wang *et al.*, 2016), and BioCreative VI Track 2: Text-mining services for Human Kinome Curation (Gobeill *et al.*, 2018). These challenges were designed for specific problems, such as finding relevant articles regarding toxicogenomics, human kinase proteins, protein-protein inter-action affected by mutations, among others.

Additionally to the works created for the BioCreAtIvE challenges, other studies have been developed describing the use of ML in the IR task to search for relevant documents. However, also none of the studies was specific for biomarkers, instead they focused on other biomedical fields, such as fungal proteins (Almeida *et al.*, 2014), HIV (Almeida *et al.*, 2016) and proteins involved in cell cycle processes (Krallinger *et al.*, 2014).

Almeida *et al.* (2014) developed a ML system for supporting the first task of the biological literature manual curation process, called triage, which involves identifying very few relevant documents among a much larger set of documents. They were looking for articles related to

characterized lignocellulose-active proteins of fungal origin to curate the mycoCLAP database (https://mycoclap.fungalgenomics.ca/). They compared the performance of various classification models, by experimenting with dataset sampling factors and a set of features, as well as three different ML algorithms (Naïve Bayes, SVM and Logistic Model Trees (LMT)). The most effective model to perform text classification on abstracts from PubMed was obtained using domain relevant features, an under-sampling technique, and the LMT algorithm, with a corresponding F-measure of 0.575.

Almeida *et al.* (2016) used a supervised learning approach to support the screening of HIV literature. They tested three different algorithms: Naïve Bayes, LMT and and SVM. The model that achieved the best results was a LMT classifier, which was trained with a training set containing 40% of "included" articles and 60% of "excluded" documents and used a Bag-Of-Words and MeSH terms as features. The classifier reached a precision of 0.467 and a recall of 0.9 for the "included" class.

Krallinger *et al.* (2014) explored the use of text mining strategies to improve the retrieval of relevant articles and individual sentences related to proteins involved in cell cycle processes. They tested their approach with abstracts and full-text articles. Classifiers built with the SVM algorithm and using the full text articles as input performed considerable better than the ones trained only with the abstracts (recall of 94% vs 57%).

2.2.2 IR in biomarker literature

The only work developed in the same TM task and domain as this thesis was created by Younesi *et al.* (2012). Despite being about improving the IR task in biomarker literature, this paper did not use ML, instead they created a dedicated biomarker terminology and performed a combined search for genes and selected classes of the biomarker retrieval terminology, hence improving the retrieval performance.

Therefore, this thesis presents the first supervised approach to improve the IR task in literature regarding biomarkers of exposure, which is likely to be more effective than terminology-based approaches, since this is the case for other IR tasks.

2.2.3 ER and RE in biomarker literature

There are several works describing the application of text-mining methods to find information about biomarkers in the literature, although none focuses specifically on biomarkers of exposure.

Lever *et al.* (2018) used a supervised learning approach to develop an Information Extraction-based method to extract sentences containing relevant relationships involving biomarkers from PubMed abstracts and Pubmed Central Open Access full text papers. With this approach, they built the CIViCmine knowledge base (http://bionlp.bcgsc.ca/civicmine/), containing over 90992 biomarkers associated with genes, drugs and cancer types. Their goal was to reduce the time needed to manually curate databases, such as the Clinical Interpretation of Variants in Cancer (CIViC) knowledgebase (https://civicdb.org/home), and to make it easier for the community curators, as well as editors, to contribute with content.

Bravo *et al.* (2014) presented a dictionary-based named entity recognition method (BioNER) and a relation extraction module to identify human gene and protein biomarkers and their associated diseases from the literature. They applied this tool on articles found on PubMed by searching with a specific query that targeted human biomarkers. The disease-related biomarkers found on these articles are publicly available at http://ibi.imim.es/biomarkers/.

Chang *et al.* (2017) created a curation pipeline to mine Hepatocellular Carcinoma (HCC) biomarkers and constructed the MarkerHub database. The methodology involves collecting abstracts from PubMed and perform ER using ML-based and pattern-based methods to identify several biological terms in the collected documents, such as genes, mutation information, cell lines and diseases related to HCC.

Summarizing, studies have been carried out to either improve the IR task using ML in other areas of the biomedical field or to perform ER and RE on biomarker data. However, none applies a supervised learning approach to the IR task on literature regarding biomarkers of exposure.

2.3 Data and text resources

2.3.1 Exposome-Explorer dataset

Exposome-Explorer (`http://exposome-explorer.iarc.fr/`) was the first database dedicated to biomarkers of exposure to environmental risk factors for diseases, created by the International Agency for Research on Cancer, part of the World Health Organization. It contains detailed information about biomarkers, such as nature, dietary and pollutant biomarkers correlation values and biomarker reproducibility values.

This project was developed using data provided by the Exposome-Explorer database curators, which includes:

1. The 9 queries used to search for citations with information about dietary, reproducibility, and pollutant biomarkers on WOS (example in Figure 2.2);

(((TI= ((pcb or pcbs or polychlorinated biphenyls) and (urine or urinary or serum or plasma or blood or adduct* or biological monitoring or level or excretion or exposure* or exposed or biomarker or biomarkers or marker or markers or metabolite or metabolites or concentration or concentrations)) and TS = (urine or urinary or serum or plasma or blood or adduct*)))) *AND* **LANGUAGE:** (English) *AND* **DOCUMENT TYPES:** (Article OR Review)
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=1900-2013

Figure 2.2: Query used to search for articles regarding the PCDD/F pollutant.

2. The WOS search results, with 8575 citations used to manually retrieve the relevant articles to curate the database. The files provided were in the BibTeX format (.bib). Figure 2.3 shows an example of a BibTeX entry;

```
@article{ ISI:000315761000006,
    Author = {Khan, Jehan A. and Jalal, Jalaluddin A. and Ioanndes, C. and Moselhy, Said S.},
    Title = {{Impact of aqueous doash extract on urinary mutagenicity in rats exposed to heterocyclic amines}},
    Journal = {{TOXICOLOGY AND INDUSTRIAL HEALTH}},
    Year = {{2013}},
    Volume = {{29}},
    Pages = {{142-148}},
    DOI = {{10.1177/0748233711427053}},
}
```

Figure 2.3: Example of a BibTeX entry present in the files with the results from the queries.

3. A text file in the TSV format with the title, first author, year, journal and PubMed ID (PMID) of the 480 relevant articles used to extract information about biomarkers of exposure for the database. From the 480 articles, 5 were not available on PubMed and therefore did not have a PMID. Figure 2.4 shows an example of this type of file, with one article per line and each attribute separated by a tab space ($\backslash t$).

title	author_first	year	journal	pmid
Folate intake assessment: validation of a new approach	Yen	2003	J Am Diet Assoc	12891147
Biomarker studies in northern Bohemia	Binkova	1996	Environ Health Perspect	8781388
Day-to-day and within-day variation in urinary iodine excretion	Rasmussen	1999	Eur J Clin Nutr	10369497
Prevalence of low chlorinated dibenzo-p-dioxin/dibenzofurans in human serum	Park	2013	Chemosphere	23062831
3-Hydroxybenzo[a]pyrene in the urine of smokers and non-smokers	Lafontaine	2006	Toxicol Lett	16406420

Figure 2.4: Example of a few lines from the file with information from the relevant publications used to build the Exposome-Explorer database.

All 480 relevant publications used to curate the database were expected to be present in the 8575 articles retrieved from the WOS, however, only 401 publications were present in the query results (Figure 2.5). The additional 79 publications used to retrieve information, were manually found by a researcher while screening the literature for relevant articles. For this reason, these scientific papers are of great value to this work and are probably even more relevant than the publications found by the query search. Chapter 4 focuses on finding a criterion to include these publications in the search results by modifying the original queries.

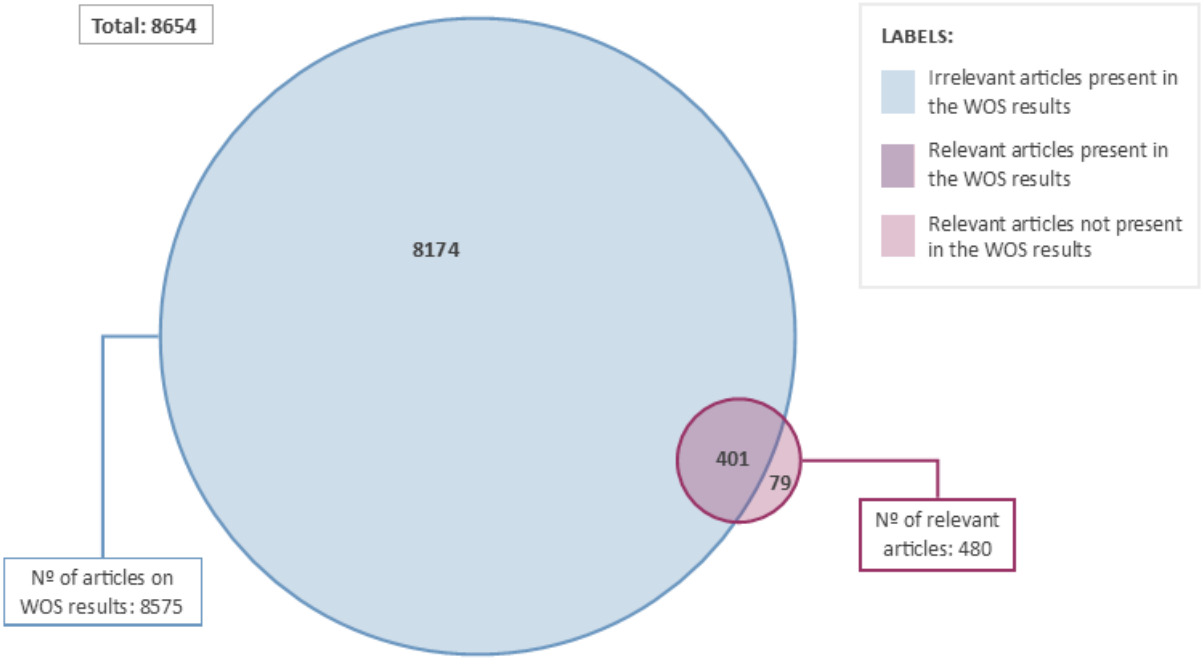

Figure 2.5: Distribution of the relevant and irrelevant articles of the WOS search results.

2.3.2 CIViCmine dataset

The CIViCmine knowledgebase (`http://bionlp.bcgsc.ca/civicmine/`) was created with TM tools to automatically extract biomarkers from literature to aid the curation of the Clinical Interpretation of Variants in Cancer (CIViC) database (`https://civicdb.org/home`). The data from the knowledgebase was downloaded from `https://github.com/jakelever/civicmine` on February 28 2019. The dataset used consists of sentences containing relevant relationships involving biomarkers, collected from abstracts.

This corpus was used to test the application of the methodology created in this thesis on other types of biomarker datasets.

2.4 Software

This project was created with Python 3.6.6. Some tasks in this project were developed with Python packages, namely NLTK, Numpy and Scikit-learn. This packages are commonly used in projects similar to this one.

Natural Language Toolkit The Natural Language Toolkit (NLTK) (Loper & Bird, 2002), is a Python package that facilitates the process of creating programs that process natural language data. NLTK is interfaced with annotated corpora and lexical resources and includes tutorials and problem sets to guide the user through computational linguistics tasks, such as classification, tokenization, and stemming. Version 3.3 was used on this project. NLTK is available under an open source license from `https://www.nltk.org/`.

NumPy NumPy (Oliphant, 2006) is a Python library that provides a multidimensional array object and a collection of fast mathematical operations on arrays, functionalities that are considered key elements for scientific programming. Python's built-in sequences are not efficient enough when applying advanced mathematical operations to large dimensional data. NumPy not only reduces the code necessary to perform such operations but also improves the speed of the program. Links to available download options at `http://scipy.org/install.html`. The version used was 1.15.4.

Scikit-learn Scikit-learn (Pedregosa *et al.*, 2011) is a Python module that allows both specialists and non-specialists to use a wide range of state-of-the-art ML algorithms, for both supervised and unsupervised learning problems. It is indicated for medium-scale data sets. This project used the version 0.20.0. Available at: `https://scikit-learn.org/`

Chapter 3

Information Retrieval using Machine Learning

This chapter focuses on using a supervised approach to create ML classifiers to predict an article's relevance for the Exposome-Explorer database in order to assist manual curation. The existing manually curated data will be preprocessed and then used to train and test the classifiers. When given a new publication, these classifiers should predict whether it is relevant to the database. The work developed on this chapter is available on GitHub (`https://github.com/lasigeBioTM/BLiR`) and submited for publication to a journal.

3.1 Methods

3.1.1 Exposome-Explorer dataset

The work on this chapter was developed using the data used to set up and develop the Exposome-Explorer database, which included:

- the WOS search results, with 8575 citations used to manually screen the relevant articles containing information about biomarkers of exposure;
- the 480 relevant articles used to extract information about biomarkers for the database.

3.1.2 Data collection

All 480 publications used to curate the database were expected to be listed in the 8575 citations retrieved from the WOS. However, only 401 of them were present: the additional 79 publications absent from the WOS query results were identified by database annotators while screening the literature for relevant articles. These 79 scientific papers were excluded from the dataset used to build the models, but will be handled on Chapter 4.

The existing dataset, listed above, was missing some important features, needed to construct the corpus for the model. For this reason, PubMed was used to extract the titles, abstracts and metadata (publication date, author names, number of times the article was cited and journal

name). The PubMed search and retrieval of PMIDs, titles, abstract and metadata was carried out with E-utilities, a public API available at NCBI Entrez system (`https://www.ncbi.nlm.nih.gov/home/develop/api/`). Some publications were found through the DOI to PMID converter and others by a combined search with the title and first author name. The resulting corpus of articles consisted of 7083 publications.

3.1.3 Data preprocessing

After retrieving the title, abstract and metadata for each article, it was necessary to prepare the text data to be used as input by the ML models (classifiers). This task included:

(i) *Assign labels to each article:* A supervised learning approach was used to build the classifiers, which means each article (document) has a known class assigned to it. To label each article, the list with the 401 articles used to curate the database was cross-referenced with the 7083 publications in the corpus. If they were present in the list, they were considered relevant and assigned the label 1. If they were not present in the list, as they were not used to extract information about biomarkers, they were considered irrelevant and therefore assigned the label 0;

(ii) *Natural Language Processing:* The text was separated into words (tokens). All words with the same root were reduced to a common form (stemming) and all stop words were removed. The tokens were then combined into n-grams;

(iii) *Transform text data to numerical data:* The ML model expects numeric data as its input. However, the titles, abstracts, and metadata are in text format. To this end, each distinct token that occurred in the documents was mapped to a numerical identifier, and the number was used to refer to the token instead of the token itself;

(iv) *Build the matrices:* Each feature represents one column and each document represents a row of the matrix. Depending on the type of matrix chosen, the matrix contained either n-gram counts (number of times each term occurs in each document) or TFIDF features (how important a n-gram is to a document in a collection of documents). An additional column was added to the training and testing data, with the respective labels. The goal of the classifier was to predict this column when applied to a new data.

The metadata from each article was handled slightly differently from the titles and the abstracts. Since it already had numerical attributes (publication date and number of citations), the matrix was created with two columns dedicated to these features, instead of having one column for each year and number of citations. The authors' names were joined into one single word (Wallace RB → WallaceRB) and were neither combined into n-grams nor went through the stemming and stop word removal stages. The journal name was preprocessed like the titles and abstracts.

Stemming was performed using the class `SnowballStemmer` from the module `nltk.stem` in the NLTK package (Loper & Bird, 2002). Steps (ii), (iii) and (iv) were carried out using the Scikit-learn (Pedregosa *et al.*, 2011) classes `CountVectorizer` and `TfidfVectorizer` from the module `sklearn.feature_extraction.text`. The main difference between the two classes is that the first one converts a collection of raw text documents to a matrix of token counts and the last one to a matrix of TFIDF features. Combinations of three different parameters were tested

to preprocess the data, resulting in different matrices used to build the classifier and, therefore, different results. The parameters tested were:

- `ngram_range (min_n, max_n)`: the lower and upper boundary of n for different n-grams to be extracted. The range of values tested were $n = \{1\}$, $n = \{1, 2\}$ and $n = \{1, 2, 3\}$;
- `min_df`: ignore all n-grams that have a document frequency lower than the given threshold. If `min_df` = 2, then terms that only appear in one article (document) will be ignored. The values of `min_df` ranged from 2 to 23, depending on the value of max_n used in the `ngram_range` parameter ($[1 + max\_n, 21 + max\_n]$);
- type of the matrix: matrix of token counts or TFIDF features.

3.1.4 Machine Learning models

The goal of the IR task was to reduce the time needed to screen the articles, by narrowing down the literature available to a set of publications that provide a reliable resource of information, in this specific case, related to biomarkers of exposure.

3.1.4.1 Building the classifiers

The ML models, also known as classifiers, were separately trained and tested using the titles, abstracts, metadata and a combination of titles + metadata, to assess which component of the article was more suitable to predict its relevance. The combination of titles and metadata was explored since the preliminary results indicated that the metadata by itself would not obtain reasonable results.

Six different algorithms were tested: Decision Tree, Logistic Regression, Naïve Bayes, Neural Networks, Random Forest and Support Vector Machine.

When given new data, the classifiers should be able to predict if a publication is relevant for the database, by labelling it with 0 for irrelevant and 1 for relevant.

The Scikit-learn package was used to test these algorithms. Most of the parameters used for each algorithm were the default ones, however, a few ones were altered to better suit the data (`class_weight, solver, kernel, gamma, bootstrap, n_estimators`), others to maximize the performance of the model (`C, alpha, max_depth, min_samples_leaf`), and one to assure a deterministic behaviour during fitting (`random_state`). The values of the parameters altered to maximize the performance of the model were found through grid search with 10-fold cross-validation. Table 3.1 summarizes the Scikit-learn functions used and the parameters changed for each algorithm.

3.1.4.2 Joining the best classifiers

When testing different classifiers using the abstracts, titles, metadata or the titles + metadata set, the prediction each model makes for a certain article might be different. The metadata model may correctly identify a publication as being relevant, while the abstracts model fails to do so. For this reason, the results of multiple classifiers were joined to understand if it was possible to retrieve more relevant publications this way. It is known that a combination of classifiers can achieve better scores than a single classifier (Dietterich, 2000; Whalen & Pandey, 2013).

Table 3.1: Scikit-learn functions and parameters for each algorithm.

	Sklearn functions	Parameters
DT	DecisionTreeClassifier	`class_weight = 'balanced'; random_state = 0; min_samples_leaf = 5`
LR	LogisticRegression	`class_weight = 'balanced'; random_state = 0; solver = 'liblinear'; C = 10.0, 1.0 or 0.1*`
NB	MultinomialNB	`alpha = 0.01`
NN	MLPClassifier	`solver = 'lbfgs'; random_state = 0`
RF	RandomForestClassifier	`class_weight = 'balanced'; random_state = 0; bootstrap = False; max_depth = 20; min_samples_leaf = 2; n_estimators = 100`
SVM	SVC	`class_weight = 'balanced'; random_state = 0; kernel = 'rbf'; gamma = 'scale'`

*$C = 0.1$ for term-count matrices; For TFIDF matrices, $C = 10.0$ for the abstracts; $C = 1.0$ for the titles and titles + metadata; $C = 0.1$ for the metadata

To try to assess whether joining the models would increase the number of relevant publications identified, the predictions from the best models for the abstracts, titles, metadata and titles + metadata were combined. Four different scenarios were tested when joining the models. In the first one, a publication was considered relevant if at least one of the models classified it as such. Then, for the other scenarios, the number of classifiers needed to consider a publication relevant was increased one by one, until it reached the total of four, in which the article had to be identified as relevant by all four models.

3.1.5 CIViCmine dataset

After building the methodology, it was important to see if it would work on another dataset that also included data about biomarkers. The dataset used was from the CIViCmine knowledgebase and consisted of sentences containing relevant relationships involving biomarkers, collected from abstracts. These sentences were considered relevant and labelled with 1. Since only the sentences used to collect information for the database were provided, it was necessary to retrieve the full abstracts in order to determine which sentences did not hold any important information for the CIViCmine database and then label them with 0 (irrelevant).

The preprocessing and IR tasks were the same used in the Exposome-Explorer dataset, also testing the combinations of preprocessing parameters and algorithms, the only difference is the fact that sentences from the abstracts are used instead of titles, full abstracts and metadata.

3.1.6 Evaluation metrics

In the data preprocessing task, labels were given to each article: 0 for irrelevant (negative) and 1 for relevant (positive). These labels were considered the gold-standard and represent the actual class of the publications.

In the IR task, all classifiers built were validated using the Scikit-learn cross-validation function (`sklearn.model_selection.cross_validate`). This model validation technique provides a

more accurate estimate of the model's performance, since it evaluates how the model will perform on new data outside the training set. The `cv` parameter of the function determines how many groups the data will be split into. In this work, a $cv = 10$ was used, which means the data was separated into 10 groups, each one was used 9 times as a training set and once as the testing set. Ten different models were built using the same parameters, with different training sets. Each time the model was fitted to the testing data, it generated a vector with predicted classes for those documents. By comparing the predictions of the testing set to the gold standard, it was possible to separate the documents into TP, FP, TN and FN.

The precision, recall and F-score were calculated for each of the ten cycles. These scores were averaged to calculate the final cross-validation metrics. Although all metrics were evaluated, the recall score was considered the most important one, since the priority of the IR task was to find as many relevant articles as possible.

By analysing the scores of the classifiers, it was possible to determine which were the best preprocessing parameters and algorithms as well as the best section of the articles (titles, abstracts, metadata or titles + metadata) to use in order to maximize the performance of the classifier.

3.2 Results

3.2.1 Exposome-Explorer dataset

3.2.1.1 Data collection and preprocessing

After data collection with PubMed, the Exposome-Explorer dataset consisted of titles, abstracts, and metadata from a total of 7083 publications. Among them, 6687 were considered irrelevant, because no information about biomarkers was extracted from them for the Exposome-Explorer database. The remaining 396 publications were considered relevant, as they were used to construct the database. The remaining 1492 articles were not present on PubMed and were excluded from the dataset.

In the beginning, all articles from all types of biomarkers in the dataset were used, however, the classifiers' metrics were very low. To try to improve the results, the data was restricted to articles regarding dietary biomarkers, since they were handled more attentively by the curators and composed the majority of the biomarkers. The new dataset consisted of 3016 publications (2860 irrelevant + 156 relevant). This attempt was successful, as the results improved.

3.2.1.2 Information Retrieval

Dietary biomarkers data A total of 120 combinations of preprocessing parameters were tested per set of abstracts, titles, metadata and titles + metadata. Since 6 different algorithms were used to build the ML models, a total of 2880 values of precision, recall and F-score were collected.

The maximum values each algorithm could reach for these metrics, using optimized parameters, are summarized in Table 3.2. For example, the maximum recall of 0.806 for the Logistic Regression algorithm on the metadata set was obtained using a `min_df` of 14, `ngram_range` (1, 1) and a term-count matrix. The parameters and algorithms used to maximize the recall for the abstracts, titles, metadata and titles + metadata can be found in Table 3.3. The highest recall score (85.8%) was obtained using the SVM algorithm on the abstracts set.

Table 3.2: Dietary biomarkers classifiers' results. Highest precision, recall and F-score reached for each algorithm per type of data.

ABSTRACTS

	Max P	Max R	Max F
DT	0.348	0.635	0.432
LR	0.632	0.743	**0.651**
NB	0.653	0.807	0.619
NN	0.718	0.571	0.580
RF	**0.834**	0.531	0.579
SVM	0.542	**0.858**	0.593

TITLES

	Max P	Max R	Max F
DT	0.224	0.624	0.312
LR	0.410	**0.774**	0.502
NB	**0.604**	0.679	**0.519**
NN	0.566	0.448	0.457
RF	0.454	0.679	0.509
SVM	0.398	0.820	0.473

METADATA

	Max P	Max R	Max F
DT	0.168	0.425	0.142
LR	0.151	**0.806**	0.245
NB	0.183	0.427	0.240
NN	0.100	0.006	0.012
RF	**0.238**	0.754	**0.263**
SVM	0.138	0.496	0.198

TITLES + METADATA

	Max P	Max R	Max F
DT	0.284	0.511	0.309
LR	0.423	**0.800**	**0.517**
NB	0.453	0.664	0.485
NN	**0.637**	0.308	0.380
RF	0.502	0.684	0.480
SVM	0.139	0.362	0.191

Joining the best classifiers To evaluate how joining the predictions from the best models for the abstracts, titles, metadata, and titles + metadata increased the recall score, four different scenarios were tested where a single publication was considered positive if at least one, two, three or four classifiers had classified it as relevant. Table 3.4 summarizes the respective results and Figure 3.1 shows how the precision and recall varied.

All biomarker publications To quantify how much restricting the dataset to dietary biomarkers had improved the results, new models were trained with the whole corpus of 7083 publications from all biomarkers using the same algorithms and parameters that had maximized the recall score for dietary biomarkers. The comparison between the values of precision, recall and F-score can be found in Table 3.5.

3.2.2 CIViCmine dataset

After data collection, a total of 7404 sentences were retrieved from 762 publications, 7059 were labelled as irrelevant and the remaining 345 as relevant. The same 120 combinations of preprocessing parameters were tested for the abstracts' sentences. Since 6 different algorithms were used to build the ML models, a total of 720 values of precision, recall and f-score were collected. The highest scores are summarized in Table 3.6.

Table 3.3: Algorithm and parameters used to get the highest recall for each set of data.

	Abstracts	Titles	Metadata	Titles + Metadata
Algorithm	SVM	SVM	LR	LR
df	21	15	14	14
n-gram	(1, 1)	(1, 3)	(1, 1)	(1, 1)
matrix	TFIDF	TFIDF	term-count	term-count
Precision	0.382	0.291	0.127	0.333
Recall	0.858	0.820	0.806	0.800
F-score	0.521	0.425	0.219	0.465

Table 3.4: Classification report for the four different scenarios of joining the results from the best classifiers. The first column represents the number of models needed for a publication to be considered positive.

	TP	FP	TN	FN	Precision	Recall	F-score
1	151	1049	1811	5	0.126	0.968	0.223
2	138	348	2512	18	0.284	0.885	0.430
3	126	206	2654	30	0.380	0.808	0.516
4	98	96	2764	58	0.505	0.628	0.560

3.3 Discussion

The highest recall score (85.8%) was obtained using the SVM algorithm on the abstracts set (Table 3.2). Among the 3016 publications used to train and test the classifier, 365 were classified as positive, which could reduce by 88% the time needed to find 85.8% of the relevant articles. Only 14.2% of the relevant articles would be lost. Looking at the results from the metadata set, globally lower values were obtained compared to other sets. This shows that the articles' metadata by itself is not informative enough to predict whether a publication is relevant for the Exposome-Explorer database. However, the abstract could be used to predict the article's relevance more effectively, similarly to how it is carried out during manual curation.

To assess whether joining the models would improve the recall score, the predictions from the best classifiers were combined (Table 3.4). When only one classifier is required to consider an article relevant, the recall score improves from 85.8% to 96.8%. However, the number of FP also increases, only reducing the number of citations to manually screen by 60.2%. When it takes two classifiers, the number of citations to manually screen is reduced by 83% and allows to find 88.5% of the relevant publications. If three classifiers are required, the number of citations to manually screen is reduced by 89% to find 80.8% of the relevant articles. At last, if a publication has to be labelled as positive by all four classifiers to be considered relevant, the number of citations to manually screen is cut down by 93.6% to find 62.8% of the relevant articles.

With this experiment it is possible to see that, for this specific dataset, a balance between all metrics must be found, however a high recall score should be prioritized, since the goal is not to miss any relevant information. If a model classified every single article as being positive, then all relevant articles would be retrieved (recall = 1.0 and precision = 0.05), however the time needed to screen the articles would stay the same. If a higher precision is favoured, the recall score lowers, fewer positive articles (FP+TP) would be retrieved, reducing the time needed to go

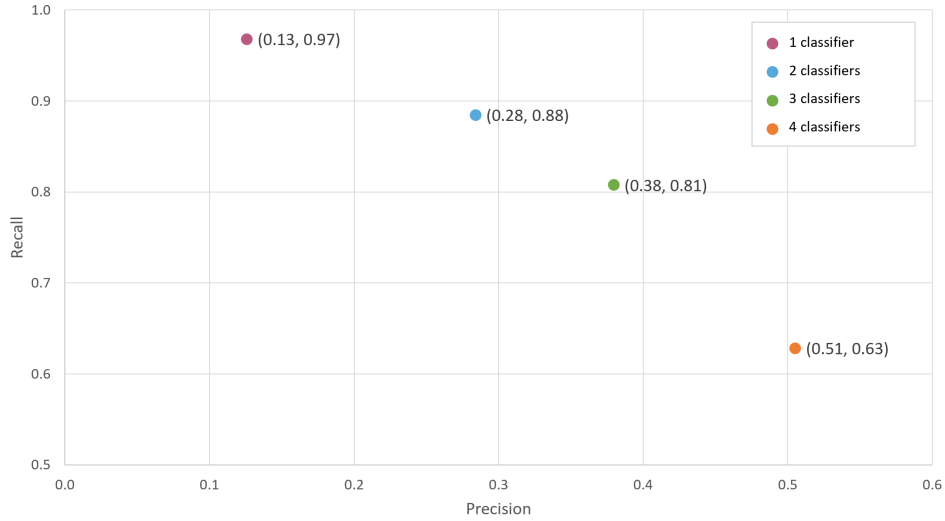

Figure 3.1: Precision vs Recall for the minimum positive classifications required for an article to be classified as relevant.

Table 3.5: Comparison of precision, recall and F-score between the dataset of all biomarkers and the dataset restricted to dietary biomarkers.

ABSTRACTS

	Precision	Recall	F-score
All	0.250	0.858	0.380
Dietary	0.382	0.858	0.521

TITLES

	Precision	Recall	F-score
All	0.224	0.795	0.343
Dietary	0.291	0.820	0.425

METADATA

	Precision	Recall	F-score
All	0.146	0.815	0.247
Dietary	0.127	0.806	0.219

TITLES + METADATA

	Precision	Recall	F-score
All	0.272	0.750	0.391
Dietary	0.333	0.800	0.465

through them, however, most relevant articles would not be found. In Figure 3.1, it is possible to see that as the minimum number of classifiers to consider an article as relevant increases, the recall decreases as the precision increases. Therefore, a lower threshold prioritizes the recall, while a higher threshold leads to a better balance between both metrics.

In order to understand why the classifiers misclassified some articles, the SVM classifier built with the titles was analysed. This classifier had a similar recall score to the abstracts one but, as the titles are shorter, they make the interpretation easier. One interesting pattern noticed was that almost all titles that had the words “food frequency questionnaire” were classified as relevant (+). From a total of 82 titles containing these words, only 2 were classified as irrelevant (both had words such as “calcium”, “water” and “energy” that were mostly found on irrelevant articles); 29 were TP and the remaining 51 were being wrongly labelled as relevant.

The title "Toenail selenium as an indicator of selenium intake among middle-aged men in an area with low soil selenium" was classified as negative, when it was in fact used in the database (FN). 39 out of 40 titles with the word "selenium" were not used in the database and thus

Table 3.6: Classifiers' results for the CIViCmine dataset.

	Max Precision	Max Recall	Max F-score
Decision Tree	0.223	0.583	0.313
Logistic Regression	0.349	0.693	0.386
Naïve Bayes	0.029	0.478	0.309
Neural Networks	**0.493**	0.327	0.360
Random Forest	0.363	0.536	**0.389**
SVM	0.378	**0.745**	0.347

labelled irrelevant: this over-represented feature may be the reason why the classifier failed to classify this article as relevant although selenium was considered of interest by the annotators.

It is also important to highlight that papers inserted in the database have been analysed considering the full-texts. This means that papers tagged as "relevant" by the classifier, could subsequently be rejected by the annotators, for a variety of reasons including "the paper is not-available online", or "the data in the paper is not presented in a way acceptable for the database". These papers would then be considered false positive by the classifier, when the title or abstract are indeed relevant.

Restricting the analysis to the dietary biomarker citations provided much better metrics than when using all the data from the database (dietary, pollutants, and reproducibility values) (Table 3.5). The reason for this improvement might be because dietary biomarkers account for almost half of the biomarkers on the Exposome-Explorer database. The focus on this type of biomarkers means they were probably handled more attentively than the others, which is reflected on the quality of the data and in the performance of the models. When restricting the analysis to citations describing the different classes of biomarkers of pollution, the performance of the models was even lower (preliminary results not shown), once again verifying the fact that the dietary biomarkers data has more quality than the other types of biomarkers and is therefore more suitable to build the classifiers.

When applying the methodology to biomarkers associated with genes, drugs and cancer types from the CIViCmine knowledgebase, the SVM algorithm had the highest recall score of 74.5%, using the parameters min_df = 2, ngram_range (1,1) and a term-count matrix. This classifier allowed to reduce the time needed to find 74.5% of the relevant sentences by 83.6%, with a loss of 25.5% of the relevant ones. This shows the methodology developed on this chapter can be applied to other types of biomarkers.

In both datasets, the algorithm with the best recall was the SVM, however it worked better on the exposure biomarkers from the Exposome-Explorer dataset. The way the CIViCmine data was collected might be the cause of such differences, since only the relevant sentences retrieved from the abstracts were provided, all the other sentences were considered irrelevant, even though they were not explicitly classified as such. The articles classified as irrelevant on the Exposome-Explorer dataset were screened to assess their relevance, however the same cannot be assured for the CIViCmine dataset.

Chapter 4

Query Improvement

This chapter focuses on improving the queries used to search for literature regarding biomarkers of exposure on the Web of Science (WOS). As previously mentioned, only 401 out of the 480 relevant publications used to curate the Exposome-Explorer database were present in the 8575 articles retrieved from the WOS. The additional 79 publications were manually found by a curator while screening the literature for relevant articles. These 79 scientific papers were excluded from the dataset used to create the classifiers in Chapter 3, as the criteria to why they were considered relevant was ill-defined. In this chapter, each excluded publication will be manually analysed to determine which alterations can be made to the queries, in order to include these articles, and other similar ones, in the WOS results.

4.1 Background

To build the Exposome-Explorer, literature regarding biomarkers of exposure was searched on the Web of Science using specific queries. There were several requirements for a scientific paper to be considered for the database: it had to be peer-reviewed, describe original work with biomarker measurements in human observational studies conducted in human populations and be available online (Neveu *et al.*, 2016). The 9 queries used and the respective results were provided by the database curators in three different categories, according to their purpose:

- **Dietary:** composed of one single query targeting records with correlation values between dietary intakes and biomarkers measured in human biospecimens;
- **Pollutant:** composed of seven different queries, each one targeting concentration values of different pollutants (Polycyclic aromatic hydrocarbons (PAH), Polychlorinated biphenyls (PCB), Polybrominated diphenyl ethers (PBDE) + Polybrominated biphenyls (PBB), Polychlorodibenzo-p-dioxins and Polychlorodibenzo-furans (PCDD/F), Heterocyclic amines (HCA), phthalates or Disinfection Byproducts (DBP)) biomarkers in human biospecimens;
- **Reproducibility:** composed of one single query targeting biomarker reproducibility values.

These queries were created specifically for the advanced search option on the WOS website, which allows to form and combine search sets. The results were restricted by language (English), document types (Article or Review) and also by timespan (July 2013, July 2014 or August 2014, depending on the set of queries). This search engine allows the usage of several query operators, such as boolean operators, proximity operators and wildcards (Table 4.1).

Table 4.1: Definition of all field tags, boolean operators, proximity operators and wildcards used in the search queries.

Field Tag	TI	Search in the title field
	TS	Search for the terms in the title, abstract, author keywords and keywords plus fields
Boolean Operator	AND	Find records containing all terms separated by the operator
	OR	Find records containing any of the terms separated by the operator
	NOT	Exclude records with certain terms from your search
Proximity Operator	NEAR/x	Find records where the terms joined by the operator are separated by a maximum number of words (x). The default value of x is 15.
Wildcard	Asterisk (*)	Any group of characters, including no character
	Question mark (?)	Any single character
	Dollar sign ($)	Zero or one character

Field tags allow to search within a record's data fields. Boolean and proximity operators are used to combine terms. Wildcards represent unknown characters and are only valid in search queries that use the English language. In the Title and Topic searches, at least three letters must succeed or precede wildcards. It is also possible to use them inside words, but not after special characters (/ @ #) and punctuation (. , : ; !). When hyphens (-) and apostrophes (') are used in names, they are treated as spaces.

4.2 Methodology overview

The dietary and reproducibility queries were run by the database curators on the WOS in July and August of 2014, respectively. The pollutant queries were run in July of 2013. When running the exact same queries in May of 2019, restricting to articles published before those dates, the number of articles retrieved is not the same as it was in 2013 and 2014 (Table 4.2), possibly because some articles were added and others removed. For example, an article was published in August 2011 but, as it was only added to the WOS Core Collection in 2017, it did not appear in the 2013 results, but will be present in the 2019 results. When testing alterations to the queries, it is important to evaluate how they differ from the original ones, in terms of number of results. However, it is not possible to know which articles were part of the WOS Core Collection in 2013 and 2014. For this reason, all comparisons between different queries were performed on the results from the search performed on May 10 2019, using the appropriate timespan.

Table 4.2: Number of articles included in the query results when running the queries in different dates. The third column corresponds to the date when the database curators run the queries. The last column corresponds to the date when the queries were run to develop this work.

Category	Date	Nº of results (date)	Nº of results (10-05-2019)
Diet	Jul 2014	3444	3732
DBP	Jul 2013	284	383
HCA	Jul 2013	157	157
PAH	Jul 2013	922	946
PCB	Jul 2013	914	929
Phthalates	Jul 2013	428	438
PBDE + PBB	Jul 2013	257	258
PCDD/F	Jul 2013	182	182
Reproducibility	Aug 2014	2140	2305

Before analysing each individual article it was necessary to determine if there was no other cause for the article not being present in the results, other than the query itself. First, each article was searched on WOS to see if it could be found: 10 out of 79 articles were not part of the WOS Core Collection, therefore it was impossible for them to be found by the search query. Then, the publication date was analysed and 6 articles were not found by the queries, not because they did not match them, but because they were published after the search was conducted. As previously mentioned, the pollutant queries were run in July 2013 and the reproducibility and dietary queries in 2014, however the Exposome-Explorer database was not created until 2016. During this time, the curators manually screened the literature and found more recent articles, adding them to the database. The remaining 63 articles were matched to the query set they belonged to: dietary, pollutant or reproducibility (Table 4.3).

Once the articles were grouped together it was easier to assess which modification could be applied to the queries in order to include as many articles as possible in the results. It was important to maintain the guidelines and keywords the curators used. For example, if a specific query condition was targeting terms related to consumption (like "intake" and "dietary"), adding the term "estimate" to that condition would not make sense. However, adding the term "ingest" would be acceptable.

Nonetheless, it is extremely important to evaluate cost-benefit of the alterations in terms of total number of articles. The goal of the search query is to narrow down the literature to a set of articles that hold important information about the targeted subject, thus, if a modification increases the number of results by thousands, it means the query will no longer prove to be beneficial for the curators, as they will have to analyse more articles that are not as specific to what they are looking for.

As seen in Table 4.3, not all pollutants have relevant articles that were not found by the queries, so this section will only focus on improving the queries that still have articles left to find (dietary, DBP, HCA, PAH, phthalates, PBDE + PBB and reproducibility). Additionally, there were 10 articles related to pollution that did not fit any of the existing pollutant chemical group queries, so a new query will also be developed to include these articles.

Table 4.3: Number of articles that were not present in the query results by reason (not part of the WOS, not found because of the date it was published and not found because it did not match the query) and by query set (dietary, pollutant and reproducibility).

Query Set	WOS	Date	Query	Total
Dietary	5	0	35	**40**
Pollutants	(4)	(5)	(20)	(29)
DBP	0	1	1	**2**
HCA	0	0	3	**3**
PAH	2	0	2	**4**
PCB	0	0	0	**0**
Phthalates	1	3	4	**8**
PBDE + PBB	0	1	1	**2**
PCDD/F	0	0	0	**0**
Other	1	0	9	**10**
Reproducibility	1	1	8	**10**
Total	**10**	**6**	**63**	**79**

4.3 Dietary biomarkers

4.3.1 Original dietary query set

The dietary biomarkers query was built with terms related to intake ("intake", "consumption", "diet", "recall", "questionnaire"), association ("association", "comparison", "correlation", "relation") and validation ("validation", "validity", "reliability", "evaluation") in order to find records with correlation values between food intake and the concentrations of biomarkers in biospecimens ("serum", "blood", "plasma", "urine", "adipose-tissue" or "hair"). The original query can be found on Figure 4.1 and the 35 articles belonging to the dietary biomarkers category are in Tables 4.4, 4.5, 4.6, 4.7 and 4.8.

TI condition #1

((TI = (((intake$ or "consumption" or "diet" or "dietary") NEAR/15 (biomarker$ or marker$ or indice$ or indicator$))

OR

TI condition #2

((intake$ or "consumption" or "diet" or "dietary") and (comparison$ or "compared" or association$ or "associated" or correlation$ or "correlated" or relation*) and (metabolite$ or concentration$ or "excretion" or "levels" or "serum" or "blood" or "plasma" or "urine" or "urinary" or "adipose-tissue" or "hair"))

OR

TI condition #3

(("evaluation" or "validity" or "validation" or "valid" or "reliability") NEAR/15 ("FFQ" or questionnaire$ or recall$ or intake$ or "consumption")))

AND

TS = ("serum" or "blood" or "plasma" or "urine" or "urinary" or "adipose-tissue" or "hair"));)

Figure 4.1: Original query used to find records with correlation values between dietary intakes and biomarkers measured in human biospecimens.

4.3.2 Alterations to the original dietary query set

The 9 articles featured in Table 4.4 could be found by the query if the words “assessing”, “assess” and “assessment” are added to the first part of the third TI condition. This alteration increases the number of results by 308.

Table 4.4: Articles with terms related to "assessment" not included in the original dietary query results.

#	PMID	Title
1	21430297	Comparison of standard methods for assessing dietary intake of benzo[a]pyrene
2	21346713	Effect of cooking loss in the assessment of vitamin intake for epidemiological data in Japan
3	19027407	Validation of a food choice map with a 3-day food record and serum values to assess folate and vitamin B-12 intake in college-aged women
4	17349090	Vitamin B6 status assessment in relation to dietary intake in high school students aged 16-18 years
5	17538531	Development of a food frequency questionnaire for the assessment of quercetin and naringenin intake
6	16293876	Assessment of a dietary questionnaire in cancer patients receiving cytotoxic chemotherapy
7	10427874	Questionnaire assessment of antioxidants and retinol intakes in Mexican women
8	8379507	The correlation between two dietary assessments of carotenoid intake and plasma carotenoid concentrations: application of a carotenoid food-composition database
9	1878353	Antioxidant vitamin intakes assessed using a food-frequency questionnaire: correlation with biochemical status in smokers and non-smokers

The articles in Table 4.5 were not found because, although the query targeted terms such as "correlation$" or "correlated" in the second TI condition, none of these terms included "correlate" and "correlates". By replacing "correlated" by "correlate$" in the query, these articles can be included in the results. This modification increases the number of results by 49.

Table 4.5: Articles with the term "correlate" or "correlates" that were not present in the original dietary query results.

#	PMID	Title
10	20576202	Twenty-four-hour urinary water-soluble vitamin levels correlate with their intakes in free-living Japanese schoolchildren
11	20502474	Twenty-four-hour urinary water-soluble vitamin levels correlate with their intakes in free-living Japanese university students
12	20417877	Urinary excretion of vitamin B1, B2, B6, niacin, pantothenic acid, folate, and vitamin C correlates with dietary intakes of free-living elderly, female Japanese

In the last TI condition, the terms "validity", "validation" and "valid" were targeted. Articles #13 and #14 in Table 4.6 were not part of the results because they have the term "validate". By adding this term to the query, only 3 new articles will be added to the results. Articles #15, #16 and #17 can be included in the results, if the terms “related”, “determinants” and “exposure” are added to the second part of the second TI condition. These adjustments increase the number of results by 177, 55 and 124, respectively.

Table 4.6: Articles related to dietary biomarkers not included in the original query results.

#	PMID	Title
13	16205743	Phyto-oestrogen intake in Scottish men: use of serum to validate a self-administered food-frequency questionnaire in older men
14	19167951	Urinary isoflavones and their metabolites validate the dietary isoflavone intakes in US adults
15	11815409	Dietary determinants of plasma enterolactone
16	8279408	Contributions of vitamin D intake and seasonal sunlight exposure to plasma 25-hydroxyvitamin D concentration in elderly women
17	2000815	Human plasma fatty acid variations and how they are related to dietary intake

The articles in Table 4.7 can be included in the results if the following TI conditions are added to the query: 1 - OR (("questionnaire" or "recall") and ("intake" or "dietary")); 2 - OR (("dietary" near/1 "intake") and ("blood" or "urinary" or "adipose tissue")))

The first new TI condition will retrieve 47 additional articles and allow articles #18, #19 and #20 to be included in the results. The second new TI condition will retrieve 183 additional articles and include articles #21, #22, #23 and #24 in the query results.

Table 4.7: Articles related to dietary biomarkers not included in the original query results that could be found by creating new TI conditions.

#	PMID	Title
18	17383269	Cruciferous vegetable intake questionnaire improves cruciferous vegetable intake estimates
19	11454500	Comparing biological measurements of vitamin C, folate, alpha-tocopherol and carotene with 24-hour dietary recall information in nonhispanic blacks and whites
20	9368807	Serum phospholipid fatty acid composition and habitual intake of marine foods registered by a semi-quantitative food frequency questionnaire
21	9752802	Dietary iodine intake and urinary iodine excretion in normal Korean adults
22	9681530	Dietary soy intake and urinary isoflavone excretion among women from a multiethnic population
23	7840075	Omega-3 fatty acids in adipose tissue of obese patients with non-insulin-dependent diabetes mellitus reflect long-term dietary intake of eicosapentaenoic and docosahexaenoic acid
24	1339083	Dietary intake of aflatoxins and the level of albumin-bound aflatoxin in peripheral blood in The Gambia, West Africa

The modifications necessary to include the articles in Table 4.8 in the query results are not viable, as including them would come with the cost of adding thousands of records to the results. For example, article #33 could be included in the results if the terms "biomarkers" and "exposure" were respectively added to the first and second part of TI condition #2. This alteration would add more than 12000 articles to the results.

Table 4.8: Articles related to dietary biomarkers not included in the original query results that could only be found using very generic expressions.

#	PMID	Title
25	20167460	Total polyphenol excretion and blood pressure in subjects at high cardiovascular risk
26	20859297	Determinants of plasma alkylresorcinol concentration in Danish postmenopausal women
27	19923368	Linoleic acid is associated with lower long-chain n-6 and n-3 fatty acids in red blood cell lipids of Canadian pregnant women
28	19022967	Frequency and type of seafood consumed influence plasma (n-3) fatty acid concentrations
29	18599176	Exploration of different methods to assess dietary acrylamide exposure in pregnant women participating in the Norwegian Mother and Child Cohort Study (MoBa)
30	17069347	Fatty acid composition of habitual omnivore and vegetarian diets
31	16923234	Home use of margarine is an important determinant of plasma trans fatty acid status: a biomarker study
32	16452910	Assessment of carotenoid status and the relation to glycaemic control in type I diabetics: a follow-up study
33	14756917	Phenolic acid metabolites as biomarkers for tea- and coffee-derived polyphenol exposure in human subjects
34	15051849	Reasonable estimates of serum vitamin E, vitamin C, and beta-cryptoxanthin are obtained with a food frequency questionnaire in older black and white adults
35	1303130	Determinants of plasma ascorbic acid in a healthy male population

4.3.3 New dietary query set

Figure 4.2 shows the improvements made to the original dietary query. The original query retrieved 3732 articles, but left out 40 relevant ones that were used in the Exposome-Explorer database to retrieve information about dietary biomarkers. The improved query has 4629 results, includes 24 more relevant articles and still leaves out 16.

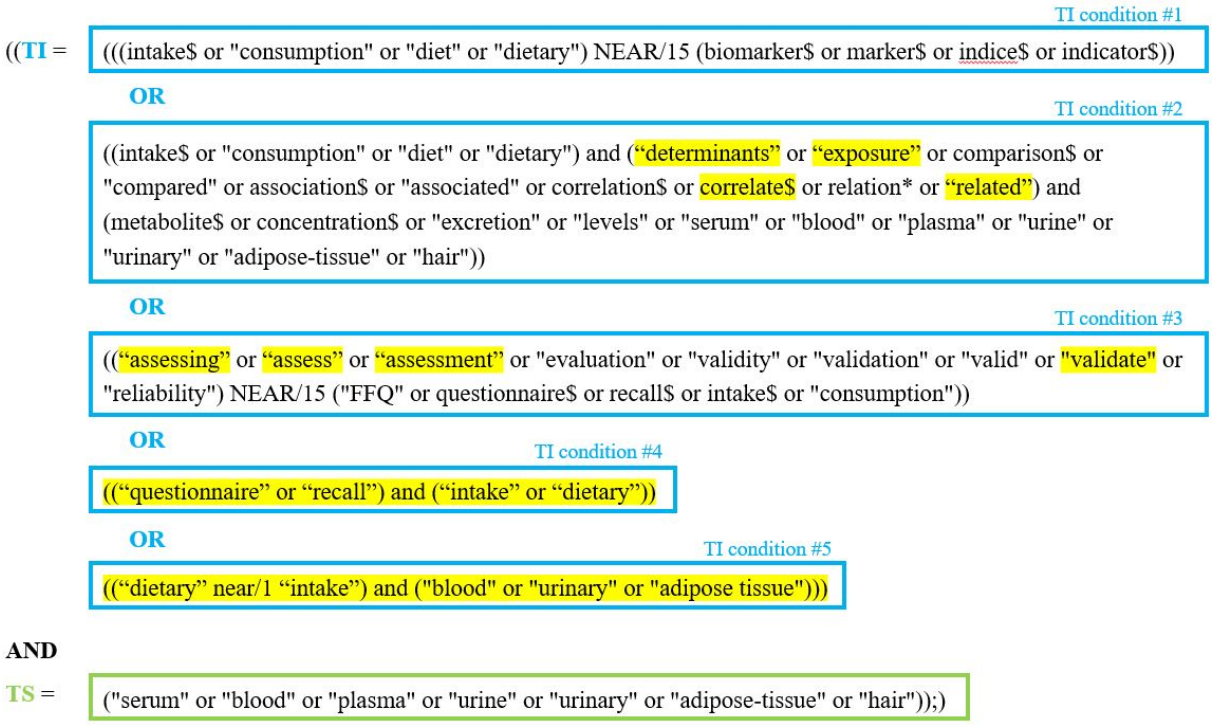

Figure 4.2: Improved dietary query. Highlighted in yellow are the alterations made to the query.

4.4 Pollutants biomarkers

4.4.1 Original pollutants query set

The pollutants set is constituted by 7 queries. They all have terms in common, such as keywords commonly associated with biomarkers ("biological monitoring", "level", "excretion", "exposure", "biomarker", "marker", "metabolite" or "concentration") and biospecimens ("urine", "urinary", "serum", "plasma", "blood" or "adduct"). They only differ in the pollutant chemical group they are targeting, which can be PAH, PCB, PBDE + PBB, PCDD/F, HCA, phthalates or DBP (Figure 4.3).

4.4.2 Alterations to the original pollutants query set

4.4.2.1 DBP

Only one article (Table 4.9) was not found by the Disinfection Byproducts (DBP) query. Although it should have been found in the results as the title has terms targeted in the TI field and the terms targeted in the TS field are present in the abstract, the abstract was not available on WOS, so the TS condition did not verify.

4.4.2.2 PAH

Articles in Table 4.10 belong to the pollutant chemical group PAH. Article #40 does not have any term commonly related to biomarkers. The word "contributors" could be added to the

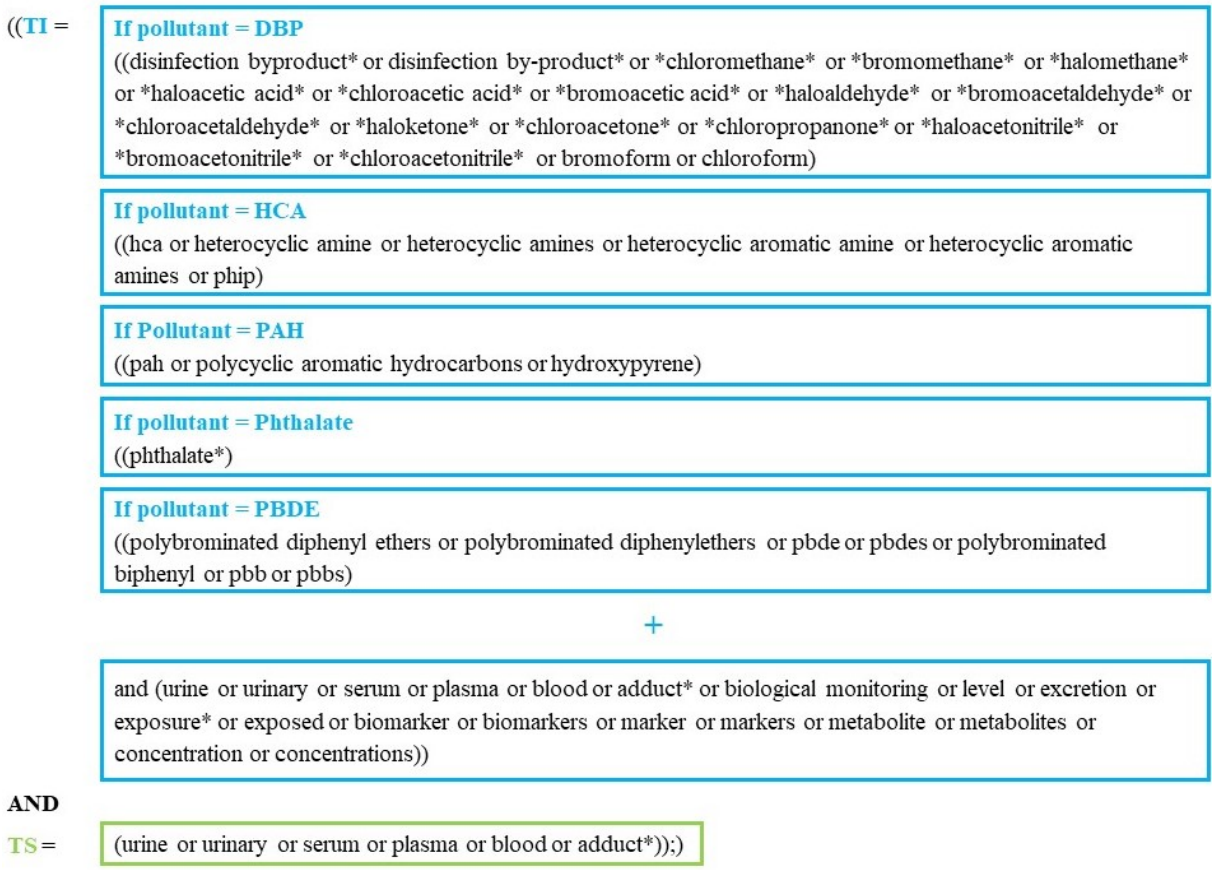

Figure 4.3: Original pollutants query set. The first part of the TI condition varies with the pollutant. The last TI condition (biomarkers and biospecimens terms) and the TS is the same for all of them.

Table 4.9: Articles related to the pollutant DBP that are not included in the original query results.

#	PMID	Title
36	17695931	Assessment of exposure of workers and swimmers to trihalomethanes in an indoor swimming pool

second part of the TI condition and the article would be found, however that word does not make sense in the context of biomarkers and would probably lead to the retrieval of other articles, that have nothing to do with biomarkers of exposure to pollutants. Article #41 was not found because the query targets the term "Hydroxypyrene" and the article has the word "3-Hydroxybenzo[a]pyrene". The article would be retrieved if "*Hydroxybenzo[a]pyrene" was added, which would increase the number of results by 11.

4.4.2.3 HCA

The HCA query targets synonyms of HCA. The three articles (Table 4.11) that belong to the this pollutant chemical group were not found by the query because they do not mention these synonyms. Article #38 has the term "2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine" which means the same as PhIP (targeted in the query). By adding this synonym to the query, the

Table 4.10: Articles related to the pollutant PAH that are not included in the original query results.

#	PMID	Title
40	19124486	Seasonal and regional contributors of 1-hydroxypyrene among children near a steel mill
41	16406420	3-Hydroxybenzo[a]pyrene in the urine of smokers and non-smokers

number of results increases by 46. As for articles #37 and #39, they mention compounds that belong to the HCA group, however they are so specific that targeting them in the query would mean specifying the search to that one case and would not retrieve other similar articles.

Table 4.11: Articles related to the pollutant HCA that are not included in the original query results.

#	PMID	Title
37	17684128	Tobacco smoking and urinary levels of 2-amino-9H-pyrido[2,3-b]indole in men of Shanghai, China
38	9270013	Urinary excretion of unmetabolized and phase II conjugates of 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine and 2-amino-3,8-dimethylimidazo[4,5-f]quinoxaline in humans: relationship to cytochrome P4501A2 and N-acetyltransferase activity
39	7920208	Urinary excretion of 2-amino-3,8-dimethylimidazo-[4,5-f]quinoxaline in white, black, and Asian men in Los Angeles County

4.4.2.4 PBDE + PBB

In order for article #42 to be found by the PBDE + PBB query, the term "breast milk" would have to be added to both the TS and TI field. This change would increase the number of articles retrieved from 258 to 311.

Table 4.12: Articles related to the pollutants PBDE + PBB that are not included in the original query results.

#	PMID	Title
42	17280703	Polybrominated diphenyl ethers (PBDEs) and polychlorinated biphenyls (PCBs) in breast milk from the Pacific Northwest

4.4.2.5 Phthalates

Three articles belong to the phthalates query (Table 4.13). Article #43 was not found because the title has the words "di-n-butylphthalate" and "butylbenzylphthalate" and the query searches terms that correspond to "phthalate*". It could be resolved by changing "phthalate*" to "*phthalate*", which would increase the results by 2. Article #44 was not retrieved because the query targets the term "biological monitoring" and this title has the term "biomonitoring". If the latest is added to the query, the results would increase by 4 articles; Article #45 and #46 have the word "DEHP" which is the same as "Bis(2-ethylhexyl) phthalate", the most common

member of the phthalates class. If this term is added to the first TI condition, the results would increase by 13.

Table 4.13: Articles related to the pollutant phthalate that are not included in the original query results.

#	PMID	Title
43	15776263	Exposure of nursery school children and their parents and teachers to di-n-butylphthalate and butylbenzylphthalate
44	23246700	Phthalate and di-(2-ethylhexyl) adipate (DEHA) intake by German infants based on the results of a duplicate diet study and biomonitoring data (INES 2)
45	15575555	DEHP metabolites in urine of children and DEHP in house dust
46	14762970	Internal exposure of nursery-school children and their parents and teachers to di(2-ethylhexyl)phthalate (DEHP)

4.4.2.6 Other pollutants

There were 9 articles (Table 4.14) that did not fit any of the pollutant queries above, because they were too specific for the pollutant chemical groups, and these articles were broader. To include them, a completely new query about pollution was created.

Table 4.14: Articles related to the general pollution that are not included in the original query results.

#	PMID	Title
47	10064554	Biomarkers for exposure to ambient air pollution–comparison of carcinogen-DNA adduct levels with other exposure markers and markers for oxidative stress
48	8781388	Biomarker studies in northern Bohemia
49	12621899	German Environmental Survey 1998 (GerES III): environmental pollutants in the urine of the German population
50	11668486	DNA adduct levels and DNA repair polymorphisms in traffic-exposed workers and a general population sample
51	8919845	Exposure to urban and rural air pollution: DNA and protein adducts and effect of glutathione-S-transferase genotype on adduct levels
52	22629390	Traffic-related air pollution and DNA damage: a longitudinal study in Taiwanese traffic conductors
53	19806728	Brominated flame retardants in serum from the general population in northern China
54	18221770	Internal exposure to pollutants measured in blood and urine of Flemish adolescents in function of area of residence
55	14564527	Urinary hydroxy-metabolites of naphthalene, phenanthrene and pyrene as markers of exposure to diesel exhaust

Like the other queries related to pollutants, two conditions needed to be present: 1) the title should target terms related to pollution, biospecimens and terms commonly associated with

biomarkers; 2) the title, abstract or keywords should mention biospecimens. All the 9 articles were individually analysed to gather terms associated with pollution ("air pollution", "traffic", "pollutants", "diesel exhaust"). As for the other terms, related to biomarkers and biospecimens, the ones used were the ones present in the other pollutant queries, however they were combined to create two conditions instead of one, in order to select less articles and increase specificity.

This query retrieved 208 articles. Articles #48, #52, #53 and #56 were not included in the results: article #48 was too general; articles #52 and #56 did not mentioned biospecimens nor biomarker-related terms; article #53 did not have any terms related to biomarkers.

4.4.3 New pollutants query set

The original pollutant query set retrieved 3293 articles, but did not include 29 relevant articles that were added by the curators. Four queries (HCA, PAH, phthalate and PBDE + PBB) from the set were improved (Figure 4.4) and a new one was created for general pollutants (Figure 4.5). The new query set retrieves 4085 results, includes 12 of these relevant articles and leaves out 17.

((TI =

If pollutant = HCA
((hca or heterocyclic amine or heterocyclic amines or heterocyclic aromatic amine or heterocyclic aromatic amines or phip or 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine)

If Pollutant = PAH
((pah or polycyclic aromatic hydrocarbons or hydroxypyrene or Hydroxybenzo[a]pyrene)

If pollutant = Phthalate
((*phthalate* or DEHP)

If pollutant = PBDE
((polybrominated diphenyl ethers or polybrominated diphenylethers or pbde or pbdes or polybrominated biphenyl or pbb or pbbs)

+

and (urine or urinary or serum or plasma or blood or adduct* or breast milk or biological monitoring or biomonitoring or level or excretion or exposure* or exposed or biomarker or biomarkers or marker or markers or metabolite or metabolites or concentration or concentrations))

AND

TS = (urine or urinary or serum or plasma or blood or adduct* or breast milk));)

Figure 4.4: Improved pollutant query set. The modification highlighted in yellow are specific to that pollutant. The ones highlighted in gray are specific to the PBDE + PBB pollutant and the one highlighted in blue is specific to the phthalate pollutant.

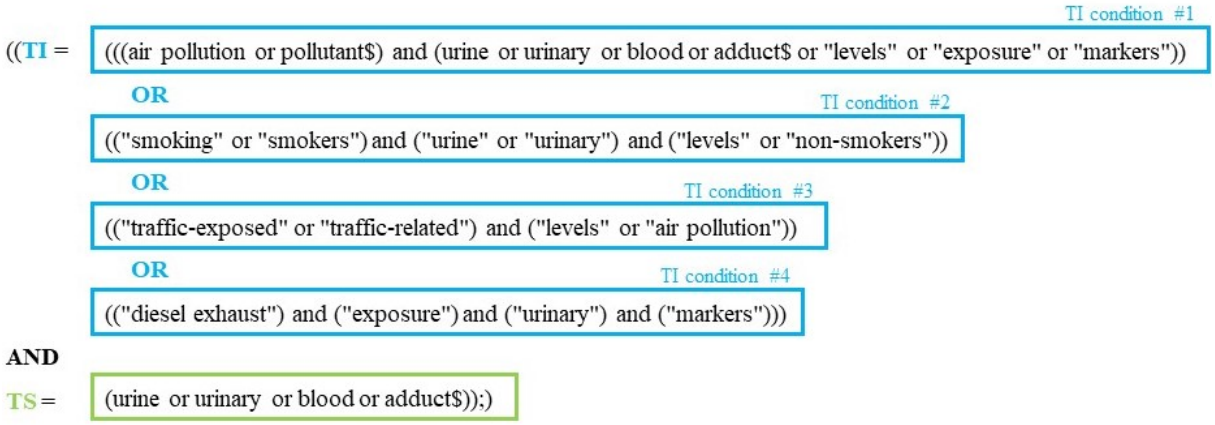

Figure 4.5: New query developed to target articles related to general pollution.

4.5 Biomarker reproducibility values

4.5.1 Original reproducibility query set

To search for biomarker reproducibility values (Figure 4.6), a combination of biomarker-related terms and reproducibility-related keywords ("variability", "reliability", "reproducibility", "repeatability", "intrasubject", "inter-subject", "within-subject", "between-subject") was used.

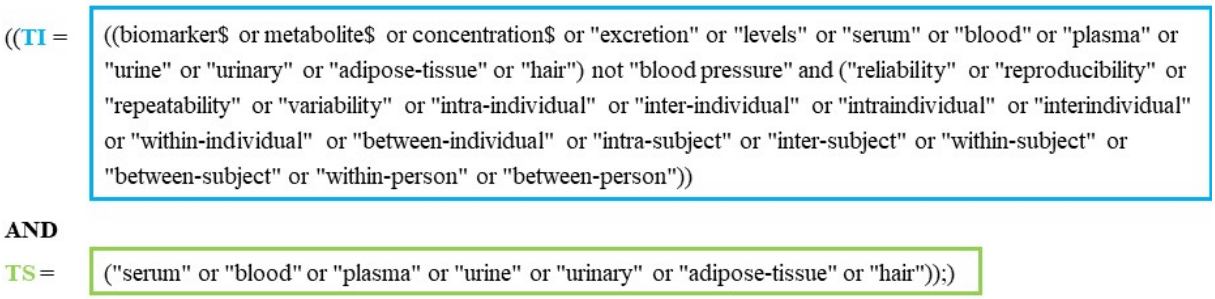

Figure 4.6: Original query used to find records with biomarker reproducibility values.

4.5.2 Alterations to the original reproducibility query set

There were seven articles related to reproducibility that were not retrieved by the query (Table 4.15). Article #56 was not found because of the last TI condition, which could be solved in three different ways: 1) By adding "day-to-day"; 2) or "within-day"; 3) or "variation". The option that adds less articles to the results is option 2), which increases the results by 44. Article #60 and #63 could be found by adding the word "metabolism" and "cross-sectional study" to the first and last TI condition, respectively, increasing the results by 83 and 465.

It was not possible to include all articles in the query results, some because the cost was greater than the benefit: for articles #57, #58 and #59 to be included, it would be necessary to add the term "variation$" to the query, which would double the results; for article #61 to be retrieved, the word "population" would need to be added, which would triple the results. Article

Table 4.15: Articles not included in the original reproducibility query results.

#	PMID	Title
56	10369497	Day-to-day and within-day variation in urinary iodine excretion
57	12507978	Short-term variations in enterolactone in serum, 24-hour urine, and spot urine and relationship with enterolactone concentrations
58	17538540	Variation in fasting and non-fasting serum enterolactone concentrations in women of the Malmö Diet and Cancer cohort
59	20332255	Long-term variation in serum 25-hydroxyvitamin D concentration among participants in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial
60	23368909	Reliability of selected antioxidants and compounds involved in one-carbon metabolism in two Dutch cohorts
61	20973429	Lead and bisphenol A concentrations in the Canadian population
62	23816546	Determinants of urinary bisphenol A concentrations in Mexican/Mexican–American pregnant women
63	16276030	Intake frequency of fish and serum levels of long-chain n-3 fatty acids: a cross-sectional study within the Japan Collaborative Cohort Study

#62 did not have any term related to reproducibility, making it difficult to include changes that respect the query structure.

4.5.3 New reproducibility query set

The improved reproducibility query (Figure 4.7) shows, the improvements made to the original query, which retrieved 2305 articles, but left out 10 relevant ones. The new query results include 3 of these relevant articles in the 2847 results but still leave out 7.

((TI = ((biomarker$ or metabolite$ or concentration$ or "excretion" or "metabolism" or "levels" or "serum" or "blood" or "plasma" or "urine" or "urinary" or "adipose-tissue" or "hair") not "blood pressure" and ("reliability" or "reproducibility" or "repeatability" or "variability" or "intra-individual" or "inter-individual" or "intraindividual" or "interindividual" or "within-individual" or "between-individual" or "intra-subject" or "inter-subject" or "within-subject" or "between-subject" or "within-person" or "between-person" or "within-day" or "cross-sectional study"))

AND

TS = ("serum" or "blood" or "plasma" or "urine" or "urinary" or "adipose-tissue" or "hair"));)

Figure 4.7: Improved reproducibility query. Highlighted in yellow are the alterations made to the query.

4.6 Discussion

Overall, the query improvements and the new query for general pollution increased the total number of articles by 2231 articles, allowing 39 more relevant articles to be retrieved (Table

Table 4.16: Comparison between the number of articles in the results from the query search on WOS and the number of relevant articles that were not found on those results before and after the query improvements.

	Nº of articles in WOS results		Nº of relevant articles not in results	
	Old Query	New Query	Old Query	New Query
Dietary	3732	4629	40	16
DBP	383	383	2	2
HCA	157	203	3	2
PAH	946	956	4	3
PCB	929	929	0	0
Phthalate	438	693	8	4
PBDE + PBB	258	311	2	1
PCDD/F	182	182	0	0
Pollution	-	428	10	5
Reproducibility	2305	2847	10	7
Total	9330	11561	79	40

4.16). It was not impossible to include 40 relevant articles in the results from the queries, because:

- 10 articles were not part of the WOS Core Collection;
- 6 articles were not found because of the timespan, which can easily be solved by running the queries again without the date restriction to July 2013 or August 2014;
- 24 articles were too different from the query structure or did not represent a cost-effective change to the query, as they increased the number of results by thousands of articles.

The query alterations proved to be the most beneficial approach to retrieve these articles and allowed the queries to be modified with explicit examples in mind, targeting a specific set of articles with similar characteristics to the ones the curators manually added to the results, that might potentially be relevant as well. This work allowed the queries to be more personalized to the curators needs, which should increase the percentage of relevant articles retrieved in the search results. Nevertheless, it was not possible to quantify the gain of altering the queries as the database curators would have to manually assess the relevance of each one of the new retrieved articles.

To evaluate how the model would perform on the new data from the improved queries, the new WOS results were downloaded and the methodology developed on Chapter 3 was applied to the data. As expected, the precision and recall of the classifiers lowered, as the quality of the data decreased from considering all those additional articles irrelevant, without them being explicitly considered as such by the database curators. Using the NB algorithm the highest recall score lowered from 0.807 to 0.751 (preliminary results not shown in this chapter).

4.7 Other approaches

Although the query improvements included almost half of the missing articles, some were still left out because of the cost of including them in the query results. Another alternative was explored to try to find a way to include all relevant articles in the results, and not just 39. The database curators shared they had found some of those articles in the references of the articles they were screening, thus, a program was developed to retrieve all references from the 401 articles found relevant for the database, to try to find the 79 articles that were not included in the results. However this approach had several issues, namely:

- Not all articles in the references are relevant, but what would be the criteria to classify them? If it was assumed the database curators had gone through all the references and ignored these articles for a reason, then the ones that were not used in the Exposome-Explorer database should be considered irrelevant, which would lower the quality of the dataset and consequently the model;
- It took more than one iteration to find the relevant articles, that is, some were only found in the references from the references of the original searched articles;
- Some of the relevant articles were only present in the references of the irrelevant articles, which meant more than 6000 article references would have to be searched to find all the relevant articles.

From the 480 articles included in the Exposome-Explorer database, 99% of these articles were found on PubMed and 98% on the WOS. From the 8174 irrelevant articles (2013/14 results) that were not included in the database, 18% were not on PubMed. Therefore, we can say that WOS has more irrelevant articles, which would suggest that PubMed could be a better resource to search for literature regarding biomarkers of exposure, instead of the WOS. Looking at the journal of the articles not present on PubMed, some have low h-index scores (< 3) and others are not even found on the Scimago Journal & Country Rank (`https://www.scimagojr.com/`), which implies these journals did not have enough impact to be considered for the MEDLINE collection. In the future, it would be interesting to adapt the WOS queries to PubMed queries (Couto, 2019) and assess if the set of articles retrieved would indeed be more restricted and relevant.

Chapter 5

Conclusions

Biomarkers of exposure are biological parameters objectively measured in the body that reflect the exposure of an individual to an environmental factor. Characterizing them is crucial for biomedical professionals to be able to correctly predict clinical responses, screen, monitor and diagnose patients, identify hazards, assess exposure and associate responses with the probability of a disease outcome. Exposome-Explorer was the first database concerning biomarkers of exposure to environmental risk factors and was developed by the International Agency for Research on Cancer, part of the World Health Organization. So far, the database has been manually constructed and more than 8500 citations were screened, but only 5% were included in the database. This method is not the most cost and time-efficient solution to either collect new data nor to keep the database updated. As the number of scientific papers continues to grow, TM techniques could be a great help to assist the triage of documents containing information about biomarkers of exposure and keep the database updated.

This thesis proposes a solution to automate the identification of relevant articles to be manually curated for the Exposome-Explorer database. There were two main objectives for this work: use a supervised learning approach to improve the IR task in order to assist the curation process of the Exposome-Explorer database and improve the original queries used by the database curators to search for articles on the WOS. All tasks were developed using data provided by the Exposome-Explorer database curators, namely, the queries used to search on the WOS for literature regarding biomarkers of exposure, the results from that search and the 480 relevant articles, selected from those results, used to construct the database.

To accomplish the first objective, PubMed was used to retrieve the titles, abstracts and metadata (journal, year, number of citations and authors) from the citations present in the WOS search results. After the data was collected it was preprocessed to create the labelled matrices to be used as input by the ML models, with rows corresponding to each article and columns to each feature. The preliminary results using data from all types of biomarkers of exposure (dietary, pollutant and reproducibility values) were not very high given the heterogeneity of the data, so they were restricted to only dietary biomarkers. With this restricted dataset, 2880 ML classifiers were created with different combinations of parameters and algorithms to access which ones provided the highest recall when predicting the relevance of articles based on their title, abstract, metadata or titles + metadata. The model with the highest recall (85.8%) was built with the SVM algorithm and used the abstracts to predict a paper's relevance. This classifier reduced the number of citations regarding dietary biomarkers to be manually screened

by the database curators by nearly 88%, while only misclassifying 14.2% of the relevant articles. When joining the results from the best classifiers for the titles, abstracts, metadata and titles + metadata, the recall score improved from 85.8% to 96.8%, when only one classifier was required to consider an article relevant. However, this value was high because the model was classifying almost all articles as relevant, thus the number of citations to manually screen was only reduced by 60.2%. The results from considering two, three and four classifiers did not outperformed the results of using only the best abstracts classifier.

This methodology can also be applied to similar biomarkers datasets or be adapted to assist the manual curation process of similar chemical or diseases databases, as was confirmed with the CIViCmine knowledgebase dataset. The methodology was adapted to predict the relevance of sentences related to biomarkers associated with genes, drugs and cancer types. The highest recall score of 74.5% was obtained with the SVM algorithm. This classifier allowed to reduce the time needed to find 74.5% of the relevant sentences by 83.6%, with a loss of 25.5% of the relevant ones.

To achieve the second objective, the original queries were improved to include 79 articles that were used to extract information about biomarkers of exposure for the Exposome-Explorer database, but that were not retrieved from the WOS search results. Overall, the query improvements and the new query for general pollution increased the total number of articles by 2231, allowing 39 additional relevant articles to be retrieved. It was still not possible to include 34 relevant articles in the results from the queries, either because they were not on WOS or because the cost of including was greater than the benefit. The remaining 6 articles were not found in 2013/14 because they were published after the curators ran the queries, which can easily be solved by running the queries again. The alterations allowed the queries to be more personalized to the curators needs, as they were modified to target specific terms that were present in real-life examples of handpicked relevant articles, which should increase the percentage of relevant articles retrieved in the search results. Nevertheless, it was not possible to quantify the gain of altering the queries as the database curators would have to manually assess the relevance of each one of the new retrieved articles.

If this methodology is incorporated into the Exposome-Explorer curation pipeline, the IR task will consist of two steps. In the first one, articles will be retrieved using the query search on WOS, to target domain-specific publications. Then, the classifier will be used to narrow down the publications even more by classifying them as relevant for the database. Manual curation will still be needed to extract information about biomarkers from full-text articles, however on a less numerous set of articles.

The main contributions of this work are a dataset, adapted from the Exposome-Explorer, with titles, abstracts and metadata from 7083 articles, classified as relevant or irrelevant, depending on whether they were used or not extract information about biomarkers of exposure and a pipeline to retrieve abstracts, titles and metadata from PubMed, preprocess them and create classifiers with a supervised learning approach. Both the data and the code are available on GitHub (`https://github.com/lasigeBioTM/BLiR`). Additionally, an article, based on Chapter 3, has also been submitted (Jesus, S., Lamurias, A., Neveu, V., Salek, R. M., & Couto, F. M. Information Retrieval using Machine Learning for biomarker curation in the Exposome-Explorer).

5.1 Future work

In the future, it would be interesting to work on improving the results from the classifiers that use the metadata set. For example, by giving different weights to the authors, according to the position they appear in, or by creating new features that result from the combinations between all authors within the same article. It would also be interesting to test different weights when joining the results from multiple classifiers. For example, the abstracts' classifier is the best one, so it should have more weight in deciding if a publication is considered relevant. When analysing why the model misclassified some publications, a few chemicals, like "calcium" and "selenium" were strongly associated with irrelevant articles. An idea to explore is to replace all chemical tokens by the same word, such as "chemical", and see if it improves the precision and recall of the classifiers, which should avoid over-fitting on the training set.

It would also be interesting to adapt the WOS queries to PubMed and assess if the set of articles retrieved would be more restricted and relevant.

In the beginning of 2019 new biomarkers were added to the Exposome-Explorer database. The new queries and the classifier could be applied to articles from 2014 until 2018 and the new data on the database could be used to further validate the methodology.

References

Almeida, H., Meurs, M.J., Kosseim, L., Butler, G. & Tsang, A. (2014). Machine learning for biomedical literature triage. *PLOS ONE*, **9**, 1–21. 9

Almeida, H., Meurs, M.J., Kosseim, L. & Tsang, A. (2016). Data sampling and supervised learning for hiv literature screening. *IEEE transactions on nanobioscience*, **15**, 354–361. 9, 10

Bishop, C.M. (2006). *Pattern recognition and machine learning*. springer. 6

Bravo, A., Cases, M., Queralt-Rosinach, N., Sanz, F. & Furlong, L. (2014). A knowledge-driven approach to extract disease-related biomarkers from the literature. *BioMed research international*, **2014**. 9, 10

Breiman, L. (2001). Random forests. *Machine learning*, **45**, 5–32. 8

Chang, N.W., Dai, H.J., Shih, Y.Y., Wu, C.Y., Rosa, D., Obena, R.P., Chen, Y.J., Hsu, W.L., Oyang, Y.J. *et al.* (2017). Biomarker identification of hepatocellular carcinoma using a methodical literature mining strategy. *Database*, **2017**. 9, 10

Couto, F.M. (2019). *Data and Text Processing for Health and Life Sciences. No. 1137 in Advances in Experimental Medicine and Biology*. Springer. 40

Cristianini, N., Shawe-Taylor, J. *et al.* (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press. 8

Dietterich, T.G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, 1–15, Springer. 17

Dietz, K., Gail, M., Krickeberg, K., Samet, J. & Tsiatis, A. (2002). Statistics for biology and health. *Survival Analysis, Edition Springer*. 7

Fausett, L.V. *et al.* (1994). *Fundamentals of neural networks: architectures, algorithms, and applications*, vol. 3. prentice-Hall Englewood Cliffs. 7

Frakes, W.B. & Baeza-Yates, R. (1992). *Information retrieval: Data structures & algorithms*, vol. 331. prentice Hall Englewood Cliffs, NJ. 7

Gobeill, J., Gaudet, P. & Ruch, P. (2018). Overview of biocreative vi kinome track. *J Biol Chem*, **744**, 7–9. 9

Hearst, M. (2003). What is text mining. *SIMS, UC Berkeley*. 5

Overview of biocreative: critical assessment of information extraction for biology. 9

REFERENCES

Jurafsky, D. & Martin, J.H. (2014). *Speech and language processing*, vol. 3. Pearson London. 6

Krallinger, M., Leitner, F. & Valencia, A. (2014). Retrieval and discovery of cell cycle literature and proteins by means of machine learning, text mining and network analysis. In *8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014)*, 285–292, Springer. 9, 10

Lever, J., Jones, M.R., Danos, A.M., Krysiak, K., Bonakdar, M., Grewal, J., Culibrk, L., Griffith, O.L., Griffith, M. & Jones, S.J. (2018). Text-mining clinically relevant cancer biomarkers for curation into the civic database. *BioRxiv*. 9, 10

Loper, E. & Bird, S. (2002). NLTK: the natural language toolkit. *arXiv preprint cs/0205028*. 13, 16

Lu, Z. & Hirschman, L. (2012). Biocuration workflows and text mining: overview of the biocreative 2012 workshop track ii. *Database*, **2012**. 9

Manning, C.D., Manning, C.D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press. 6

Matis-Mitchell, S., Roberts, P., Tudor, C.O. & Arighi, C.N. (2013). Biocreative iv interactive task. In *Fourth BioCreative Challenge Evaluation Workshop. Bethesda, MD*, 190–203. 9

Neveu, V., Moussy, A., Rouaix, H., Wedekind, R., Pon, A., Knox, C., Wishart, D.S. & Scalbert, A. (2016). Exposome-explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. *Nucleic Acids Research*, **45**, D979–D984. 1, 25

Oliphant, T.E. (2006). *A guide to NumPy*, vol. 1. Trelgol Publishing USA. 13

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, **12**, 2825−2830. 13, 16

Quinlan, J.R. (1986). Induction of decision trees. *Machine learning*, **1**, 81–106. 7

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, **34**, 1–47. 3

Singh, P.K. & Husain, M.S. (2014). Books reviews using naıve bayes and clustering classifier. In *Conference: Second International Conference on Emerging Research in Computing, Information, Communication and Applications'(ERCICA-14)*, 886–891. 7

Strimbu, K. & Tavel, J.A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS*, **5**, 463–466. 1

Wang, Q., S Abdul, S., Almeida, L., Ananiadou, S., Balderas-Martínez, Y.I., Batista-Navarro, R., Campos, D., Chilton, L., Chou, H.J., Contreras, G. *et al.* (2016). Overview of the interactive task in biocreative v. *Database*, **2016**. 9

Whalen, S. & Pandey, G. (2013). A comparative analysis of ensemble classifiers: case studies in genomics. In *2013 IEEE 13th International Conference on Data Mining*, 807–816, IEEE. 17

WHO, I. (1993). Environmental health criteria 155, biomarkers and risk assessment: Concept and principles. *World Health Organization, Geneva*. 1

Witten, I.H., Frank, E., Hall, M.A. & Pal, C.J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann. 3

Younesi, E., Toldo, L., Müller, B., Friedrich, C.M., Novac, N., Scheer, A., Hofmann-Apitius, M. & Fluck, J. (2012). Mining biomarker information in biomedical literature. *BMC medical informatics and decision making*, **12**, 148. 9, 10